UNIVERSITY OF ALBERTA

RECOVERY OF DECISION CONSISTENCY IN SHORTENED AND MODIFIED SCHOLARSHIP EXAMINATIONS

by

Don Albert Klinger ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Educational Psychology

Edmonton, Alberta

Fall, 2000

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-59613-3

Canada

University Of Alberta

Library Release Form

Name of Author: DON ALBERT KLINGER

Title of Thesis: RECOVERY OF DECISION CONSISTENCY IN SHORTENED AND MODIFIED SCHOLARSHIP EXAMINATIONS

Degree: DOCTOR OF PHILOSOPHY

Year this Degree Granted: 2000

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

423RH Michener Park,
Edmonton, Alberta.
T6H 4M5

August 29, 2000

University Of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the faculty of graduate Studies and Research for acceptance, a thesis entitled RECOVERY OF DECISION CONSISTENCY IN SHORTENED AND MODIFIED SCHOLARSHIP EXAMINATIONS submitted by DON ALBERT KLINGER in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Rogers
Dr. W. Todd Rogers

Dr. Mark J. Gierl

Dr. Xin Ma

RH Short
Dr. Robert H. Short

Dr. Marcel Bouffard

Dr. Phillip Nagy

Date: August 28, 2000

Abstract

The British Columbia Ministry of Education recently modified its scholarship examination procedure by removing a substantial portion of the examinations used to determine scholarship recipients from the population of graduating High-School students. The primary purposes of this study were to 1) examine the effects of the change in the procedure used to determine scholarship recipients and 2) determine if alternative, cost effective procedures could be used to better replicate the original scholarship decisions in comparison to the current procedure. The use of the alternative procedures also provided an opportunity to examine if hypothesized interactions and differences occurred between procedures.

Correlations, root mean square error analyses, and decision consistencies were examined based on the Biology, Chemistry, Geography, Geology, Math, and Physics examinations written in January or June of 1994/95 and 1995/96. These examinations contained both dichotomously-scored multiple-choice items and polytomously-scored extended-response items. Along with the current procedure, 17 alternative procedures were examined that incorporated the generalized partial credit model (GPCM), differential weighting of the multiple-choice and the extended-response sections, and/or auxiliary information.

The use of the current procedure produced a 10% error rate in the scholarship decisions as compared to the original procedure. The majority of the wrong decisions were false negative decisions, those that would deny a student a scholarship using the current procedure although the student received a scholarship using the original procedure. Alternative procedures were unable to improve upon this error rate indicating that these procedures were randomly equivalent in terms of estimating achievement to the total test score model currently in use. Further, the simultaneous or separate estimation of the multiple-choice and extended-response items using the GPCM procedure in PARSCALE 3.1 produced very similar results suggesting that these item formats were measuring an essentially unidimensional trait. Estimation problems were noted with the use of the three-parameter model in PARSCALE 3.1. Previous findings that females outperform males on extended response items and that extended response items are better predictors of achievement for high ability students were not supported by this study. Policy implications and directions for future research are provided.

Acknowledgements

I would like to take this opportunity to acknowledge and thank those individuals who have supported and helped me achieve this goal. I can assure you that I would not be writing this if it were not for all of you. First, and most importantly, I would like to thank my beautiful wife, Marsha Gwendolyn Klinger, not only for the sacrifices she has made in order to support me in this endeavour but also for her continued love, encouragement, guidance, and help. It was your courage and fortitude that enabled me to pursue and complete my studies. I hope I can do the same for you. I also would like to acknowledge my daughters, Leisha Elizabeth and Shayla Megan Klinger, who had little choice but to come up to Edmonton while "dad went back to school." They left their friends for an adventure that I hope they will be able to cherish.

I would like to express my gratitude to our family and friends who are too many to name but too important not to acknowledge. In particular, my parents, Henry and Betty Klinger, for always being willing to offer their help in any way that they could. My desire to learn is largely due to the two of you. Next, I can honestly say that I never would have attempted let alone accomplished this goal without the support of Marsha's parents, Annamarie and the late John Hamilton Mathers. You stressed the importance of further education and supported us with all of your hearts. Much of my decision to pursue my education is due to your support and I do not think it is possible to completely show my gratitude to both of you. Thus, to you Annamarie and in fond remembrance of John, thank you.

Next, I would like to thank the teachers and professors, both past and present, that have taught and inspired me. Specifically, W. Todd Rogers, who continually encouraged me to pursue what I did not believe was in my future. This accomplishment is largely the responsibility of you and your family. You gave me with the opportunities and support I needed to complete this goal and enabled me to pursue other avenues of research. I hope I will be able to provide others the same guidance you have provided me. Mark Gierl taught me concepts that I had previously never encountered and more importantly, engaged me in conversations and discussions that made me carefully examine my own preconceptions and thoughts about learning and assessment. I want to stress how much you continually challenged me academically. Xin Ma worked with me to explore research methodologies that will let me examine the practical aspects of educational research I have always wanted to pursue. There are many others as well, of which David Bateson, Carl Walters, and Fred John foremost come to my mind.

Lastly, I would like to acknowledge and thank the students in CRAME collectively who have helped to make it a great place to work and learn. I have thoroughly enjoyed my experiences at the University of Alberta and much of it is due to all of you. Specifically, Tess (Teresa) Dawber whose friendship and competitive spirit have been a great source of inspiration; you have and will always be a great friend. Keith Boughton, with his work ethic, has made our working and research relationship both enjoyable and productive. Finally, my acknowledgements would be incomplete if I ended without mention of Michael Jodoin and Dianne Henderson. The two of you were much more than fellow students, “All for One and One for All!” I will always remember our times together and I look forward to our future encounters as colleagues and friends. If they are anything like the ones we have already shared, they will be memorable.

Table of Contents

CHAPTER 1 ... 1

Purpose ... 5
Rationale ... 7
Definition of Terms ... 8
Delimitations ... 9

CHAPTER 2 REVIEW OF THE LITERATURE ... 11

The British Columbia Model ... 11
  Purpose of the Provincial Examinations ... 12
  Purpose of the Scholarship Examinations ... 13
  Construction of the Provincial Examinations ... 13
  Construction of the Scholarship Examinations ... 17
  Administration of the Provincial Examinations ... 17
  Administration of the Scholarship Examinations ... 18
  Scoring of the Provincial Examinations ... 18
  Scoring of the Scholarship Examinations ... 21
  Reporting of the Provincial Examination Results ... 21
  Reporting of the Scholarship Results ... 21
  Potential Problems with the Current Procedure ... 23
Psychometric Models ... 26
  Classical Test Score Theory ... 26
  Item Response Theory (IRT) ... 27
  Dichotomous Item Response Models ... 28
  Polytomous Item Response Models ... 31
A Comparison of Classical Test Score Theory and Item Response Theory ... 34
The Use of Item Response Theory ... 37
  IRT Assumptions ... 38
The Value of Extended-Response Items ... 42
The Combined Use of Multiple-Choice (MC) and Extended-Response (ER) Items ... 44
Subtest Weighting and Auxiliary Information ... 46
  Auxiliary Information ... 48

CHAPTER 3 METHODS ... 50

The Data Set ... 50
Stage 1: Data Integrity ... 52
Stage 2: Scholarship Score Calculation Using the Current Procedure ... 53
Stage 3: Examination of the Accuracy of the Current Procedure ... 54
Stage 4: Scholarship Score Calculation Using the GPCM ... 59
  Underlying Assumptions of IRT ... 59
  Calculation of the Scholarship Scores ... 62
Stage 5: Examination of the Accuracy of the GPCM ... 63
Stage 6: Scholarship Score Determination using Subtest Weighting ... 63
Stage 7: Scholarship Score Determination using Auxiliary Information ... 65

CHAPTER 4 ANALYSIS OF THE CURRENT PROCEDURE 68

SECTION 1: DATA INTEGRITY 69
SECTION 2: SCHOLARSHIP SCORE CALCULATION USING THE CURRENT PROCEDURE 69
SECTION 3: EXAMINATION OF THE ACCURACY OF THE CURRENT PROCEDURE 72
Individual Examination Scholarship Scores 73
Total Scholarship Score Results 80
Decision Consistency in Subsets of Examinations 83
Decision Consistency for Males and Females 87

CHAPTER 5 ANALYSIS OF THE ALTERNATIVE PROCEDURES 89

MODEL/DATA FIT FOR THE IRT MODELS USED IN THE STUDY 89
THE USE AND EFFECTIVENESS OF THE ALTERNATIVE PROCEDURES 94
Scholarship Decision Consistency Using the GPCM 99
Scholarship Decision Consistency Using Subtest Weighting 99
Scholarship Decision Consistency using School-Based Mark as Auxiliary Information 101
Summary 101
The Effect of Gender on Decision Consistency for the Alternative Procedures 102
PSYCHOMETRIC ISSUES ASSOCIATED WITH THE ALTERNATIVE PROCEDURES 106
The Effects of Model Data Fit on Ability Estimation 106
The Differences Between Simultaneous Estimation, Separate Estimation, and Weighting of the MC and ER Sections 107
The Use of School-Based Mark as Auxiliary Information 111
The 3-Parameter Model in PARSCALE 3.1 113

CHAPTER 6 SUMMARY AND CONCLUSIONS 116

SUMMARY OF RESEARCH QUESTIONS AND METHODS 116
FINDINGS 118
Comparison of the Original and Current Procedures 118
Comparison of the Alternative Procedures to the Original and Current Procedures 119
Psychometric Issues 120
LIMITATIONS 121
CONCLUSIONS AND IMPLICATIONS FOR PRACTICE 123
FUTURE RESEARCH 130

REFERENCES 135

APPENDIX A 144

COMPARISON OF THE EXAMINATION SCHOLARSHIP RESULTS BETWEEN THE ORIGINAL AND THE ALTERNATIVE PROCEDURES 144

List of Tables

Table 1 Value, Format, and Description of the Provincial Examinations used in the Present Study 16
Table 2 Number of Students Writing Provincial and Scholarship Examinations 50
Table 3 Sample Confusion Matrix for Examination Scholarship Scores Classification Consistency 56
Table 4 Sample Confusion Matrix For Total Scholarship Score Decision Consistency 57
Table 5 Guidelines for School-Based Mark (SBM) Transformation 65
Table 6 Descriptive Statistics for Scholarship Scores using the Original and Current Procedures 70
Table 7 Comparison of Scholarship Results Between the Original and Current Procedures 74
Table 8 Confusion Matrix for the Biology 12 June 1995 Examination 77
Table 9 Scholarship Examination Classification Errors using the Current Procedure ... 78
Table 10 Confusion Matrix for the Scholarship Decisions Comparing the Original and the Current Procedures 82
Table 11 Scholarship Decision Errors using the Current Procedure for Subsets of Examinations 84
Table 12 The First 5 Eigenvalues (EV) and the Proportion of Variance Accounted for by the First Factor for the Provincial Examinations 91
Table 13 *P*-values for the three Most Difficult Multiple-Choice Items on Each Examination for the Full sample and the Lowest Decile of Students 95
Table 14 Error Rates for the Scholarship Decisions using the Alternative Procedures.... 98

Table 15 Error Rates for the Scholarship Decisions Based on Gender using the Alternative Procedures for 1994/95 103

Table 16 Error Rates for the Scholarship Decisions Based on Gender using the Alternative Procedures for 1995/96 104

Table 17 Descriptive Statistics for Item Parameters for Different Estimation Procedures 109

Table 18 Item Parameter Correlations between Different Estimation Procedures 110

Table 19 Comparison of Examination Scholarship Results Between the Original and the GPCM Procedures 145

Table 20 Comparison of Examination Scholarship Results Between the Original and the Classical, (Scholarship) Procedures 146

Table 21 Comparison of Examination Scholarship Results Between the Original and the Classical, (Optimal) Procedures 147

Table 22 Comparison of Examination Scholarship Results Between the Original and the GPCM, (2-parm; Scholarship) Procedures 148

Table 23 Comparison of Examination Scholarship Results Between the Original and the GPCM, (2-parm; Optimal) Procedures 149

Table 24 Comparison of Examination Scholarship Results Between the Original and the GPCM, (3-Parm; Scholarship) Procedures 150

Table 25 Comparison of Examination Scholarship Results Between the Original and the GPCM, (3-Parm; Optimal) Procedures 151

Table 26 Comparison of Examination Scholarship Results Between the Original and the Classical, MC-ER-SBM Procedures 152

Table 27 Comparison of Examination Scholarship Results Between the Original and the Classical, and SBM (Optimal) Procedures 153

Table 28 Comparison of Examination Scholarship Results Between the Original and the Classical, MC/ER/SBM (Optimal) Procedures........ 154

Table 29 Comparison of Examination Scholarship Results Between the Original and the GPCM, MC-ER-SBM Procedures........ 155

Table 30 Comparison of Examination Scholarship Results Between the Original and the GPCM, MC-ER/SBM (0.90, 0.1) Procedures........ 156

Table 31 Comparison of Examination Scholarship Results Between the Original and the GPCM, MC-ER/SBM (Optimal) Procedures 157

Table 32 Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER/SBM (0.45, 0.45, 0.1) Procedures........ 158

Table 33 Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER/SBM (Optimal) Procedures........ 159

Table 34 Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER-SBM (0.5, 0.5) Procedures 160

Table 35 Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER-SBM (Optimal) Procedures 161

List of Figures

Figure 1. Comparison of the Scholarship Scores Calculated using the Original and the Current Procedures 24

Figure 2. Item Response Curves for Two Dichotomously-Scored Items 30

Figure 3. Item Category Response Curves for a Four-category Item 33

Figure 4. Flow Chart Illustrating the Procedures used and the Comparisons Made........ 67

Figure 5. Students with Scholarship Scores from either the Original or the Current Procedures 72

Figure 6. Comparison of Scree Plots for the June 1995 Biology and Geography Examinations 90

Figure 7. Scholarship Score Distributions for the GPCM Procedure 96

CHAPTER 1

The use of large-scale testing is varied in terms of purpose, function, and analysis. Twenty-two states in the United States use state-wide testing as part of High-School students' graduation requirements (Council of Chief State School Officers, 1998). In Canada, British Columbia, Alberta, Manitoba, Quebec, New Brunswick, and Newfoundland have provincial examination programs that help determine High-School students' grades and Ontario is considering the implementation of a similar program (Cheliminsky & York, 1994; Lafleur & Ireland, 1999). Such examination programs are considered high stakes because of the implications of the results to the students. Historically, such examinations have been based on classical test score theory (CTST), using a single mathematical model relating a student's total test score to achievement.

Lord (1952) proposed an alternative to classical test score theory that combines student performance with item characteristics, for example, item difficulty and discrimination, to determine a student's level of achievement. This modern test score theory, now commonly called item response theory (IRT), uses a series of mathematical models that proponents claim have several advantages over the classical test score theory framework. Chief among these advantages is the ability to select different models enabling psychometricians to choose the model that best fits the characteristics of the student responses as well as the examination items. The item characteristics in combination with the response pattern of each student are used to provide estimates of ability (θ). Thus, two students with the same raw score but different response patterns could receive different θ estimates. The student who correctly answered the more difficult and discriminating test items would receive a higher score (θ estimate) than the student who correctly answered the same number of easier and less discriminating items. Although computationally complex, the increase in computing power increasingly is enabling psychometricians to use IRT and its associated mathematical models (functions) as an alternative and perhaps more accurate method to estimate achievement than the estimations produced using total test score. Currently, large testing companies use IRT as the foundation for measuring achievement with examinations having either multiple-choice (MC) or extended-response (ER) items. In contrast, despite the apparent advantages of IRT and the availability of computer programs that can quickly complete the analyses, state and provincial testing officials continue to rely on classical test score theory. Given the consequences of many state and provincial examination programs, it is essential that the results provide an accurate measure of achievement or student proficiency. Therefore, it is important to investigate examination programs with respect to their accuracy of measurement and to determine if either the classical or IRT models provide superior results.

Due to recent changes in its examination program, the British Columbia provincial examination system provides an opportunity to compare the results reported using either the classical or IRT framework within a high stakes examination system. A unique program in North America, British Columbia High-School students complete curriculum based examinations in their academic classes, with the results being used to help determine student grades and award provincial academic scholarships to high achieving students. In particular, the provincial scholarship program is of interest in the current study. Beginning in 1974 and continuing through 1983, High-School students who were interested in obtaining a provincial academic scholarship wrote optional two-hour scholarship examinations in their grade 12 academic courses. For each course, an examination was developed containing curriculum based but somewhat difficult ER items. Students who achieved a minimum stanine of five on each of their three highest scholarship examinations and an average stanine of seven or higher on these three examinations received a scholarship.

Two significant changes occurred in 1984. First, the provincial government reintroduced mandatory provincial examinations for all grade 12 academic courses. For each academic grade 12 provincial course, a student's final grade was to be based on the school-based mark (SBM) and the mark obtained on a two-hour curriculum based provincial examination. Each of the examinations consisted of a set of MC and ER items that encompassed the major concepts within each curriculum. Second, the provincial government melded the scholarship program together with the new provincial examination program. Along with completing a provincial examination for a specific course, students interested in obtaining a provincial scholarship also wrote the optional scholarship examination. As with the previous scholarship program, the new scholarship examinations consisted of conceptually difficult ER items but the length of the examination was reduced to one hour. In order to compensate for the shorter scholarship examination, an individual student's scholarship score was calculated using a simple sum of the provincial examination score (not including SBM) and the scholarship examination score. Thus, students interested in receiving a scholarship would have three hours of testing. Due to the length of the two examinations within each course, the total score value of each scholarship examination was one-half the total score value of the corresponding provincial examination.

As before, students who wrote the scholarship examinations had to meet the scholarship score standards based on their highest three scholarship scores in order to receive a scholarship. However, the scoring system used to determine scholarship recipients was different. Within each course, the raw scholarship scores, obtained by adding the provincial and scholarship examination scores of the students who wrote both components, were normalized using a ranking procedure and then transformed onto a standardized score scale having a mean of 500 with a standard deviation of 100, with the minimum and maximum scores set to 200 and 800, respectively. The cutoff points used to identify scholarship recipients were also modified to reflect these changes. First, the minimum scholarship score required for each of the three highest examinations was changed from a stanine of five to a scaled score of 475. Second, the minimum required total scholarship score over the three highest examinations was changed from an average stanine of seven to a minimum total scholarship score of 1700.

Beginning with the 1996/97 school year, the provincial government eliminated the scholarship examination but maintained the scholarship program itself. One reason for the elimination of the scholarship examinations was the cost of the development and marking of these examinations during a period of time when government expenditures were being reduced and departments were asked to find ways to reduce costs. The belief was that the information gained from the scholarship examinations was redundant with the information from the provincial examinations because the scores from both examinations were highly correlated (Ron West, personal communication, September 22, 1999). Given this change, scholarship scores are now based solely on the provincial examinations. However, since the scholarship examinations were optional and only a portion of the population of students enrolled in academic grade 12 courses attempted to obtain a provincial scholarship, the calculation of scholarship scores are now based on only a portion of the students writing each provincial examination and only this portion of students are given a scholarship score. For each course, this is accomplished by including only those students having a provincial examination score of 70% or higher. Within this sub-sample, the scores are normalized using a ranking procedure and scaled so that the average is 500 and the standard deviation is 100, with the minimum and maximum scores being 200 and 800. As before, scholarships are awarded to students obtaining a scholarship score of at least 475 on their three highest examinations and having a combined minimum total of 1700 based on these three scores.

Purpose

The implications of the most recent changes in the scholarship examination program in British Columbia have not been fully assessed. What influence, if any, upon the identification of scholarship winners was introduced because of the discontinuation of the optional one-hour scholarship examinations? Since there has not been a change in the difficulty of the provincial examinations, the examinations used to determine scholarship recipients are now shorter and simpler. This may be problematic because the scholarship examinations that were designed to increase the dependability of scholarship decisions are no longer present. Further, when the scholarship examinations were in place, students having the same provincial examination score could have different scholarship scores because of differential performance on the scholarship examinations. With the current procedure, this differentiation is eliminated since a given provincial examination score is now associated with a single scholarship score. Consequently, the questions addressed in the present study were:

1. How has the elimination of the one-hour scholarship examinations changed which students receive scholarships?

2. Can the use of alternative approaches incorporating item response theory in the form of the Generalized Partial Credit Model (GPCM), weighting of the MC and ER sections, and/or including auxiliary information in the form of school-based mark improve upon the decisions made?

In addition, an ancillary question was addressed:

1. Do theorized interactions and differences occur when alternative approaches are compared?

Overview of the procedure. To address these questions, the study was completed in seven sequential stages. Analyses were based on the last two school years in which the one-hour scholarship examinations were written, 1994/95 and 1995/96. The original scholarship scores that were obtained by students based on both the provincial and scholarship examinations were considered the 'gold standard' upon which the procedures using only the provincial examination were compared. In the first three stages, individual student scholarship scores were calculated using the current procedure. The results of these three stages were then used to answer question one above and determine the extent of the differences in scholarship scores and decisions due to the change in policy. The remaining four stages were used to investigate the second question above. Stages four and five were used to calculate the scholarship scores using the generalized partial credit model (GPCM). The final two stages were used to examine the benefit of MC and ER weighting and the use of auxiliary information used with not only the current procedure but also the GPCM. At the same time, given the limited research on the GPCM and the other alternative procedures, the results of these four stages were also used to address the ancillary question for examinations containing both MC and ER items.

Rationale

Both policy and psychometric issues were addressed in the present study. From a policy perspective, the removal of the scholarship examinations may change not only the number of provincial scholarships awarded, but also the students who receive provincial scholarships. Such changes may unfairly harm or benefit students who are attempting to qualify for scholarships. Given the decision to eliminate the scholarship examinations was largely a financial decision, it is important to examine if using alternative, readily available, and cost-effective estimation methods possibly in combination with weighting and auxiliary information, can better replicate the original results than the current estimation procedure.

From a psychometric perspective, results from CTST and IRT were compared with the original scholarship data. Over the past decade, there has been an increased use of performance and ER items in large-scale achievement tests (e.g., Council of Chief State School Officers, 1998). Yet, the examination of the utility of IRT models for polytomous data have not been carefully examined under actual conditions. It is important to determine if the models within IRT are advantageous in terms of estimating student achievement. Currently, practitioner unfamiliarity with such models and the notion that students with identical raw test scores can obtain different ability estimates has largely limited the use of IRT to only the most sophisticated examination programs.

Evidence that IRT models provide better estimates would help to justify their use in high-stakes examination programs. If such advantages were realized, then research would shift to focus on issues of increased implementation of these models and IRT in general.

Definition of Terms

Auxiliary information. Auxiliary information is additional information or data derived from sources outside the examination that is used to improve the estimation of either item parameters or ability estimates. For example, student grade-point averages, school-based mark, or previous test scores could all be used as auxiliary information.

Current procedure. The current procedure now being used to calculate examination scholarship scores uses only the provincial examination scores to determine scholarship scores for the subset of students who have a minimum score of 70% on a specific provincial examination.

Dichotomously-scored item. A dichotomously-scored item is an item that is scored on a two-point scale, typically one for a correct response or best answer and zero for all other responses. Multiple-choice items are an example of such items.

Examination scholarship score. An examination scholarship score is the scholarship score that a qualifying student receives based on the examination results for a single course. Examination scholarship scores are reported on a score scale having a mean of 500 and a standard deviation of 100.

Original procedure. The original procedure used to calculate examination scholarship scores was based on the summed scores of both the provincial and the optional scholarship examinations. Students who wrote the optional scholarship examination would be given a scholarship score for each subject area provincial/scholarship examination written. These scores were reported on a score scale having a mean of 500 and a standard deviation of 100.

Polytomously-scored item. A polytomously-scored item is an item having three or more ordered score points (e.g., 1, 2, and 3). Essays, restricted essays, mathematical problems, and performance tasks are examples of items that typically require polytomous scoring.

Total scholarship score. The total scholarship score for each student is the summed score of that student's three highest examination scholarship scores. Only those examination scholarship scores at 475 or above are included in the total scholarship score for each student. If the total scholarship score is at least 1700, the student is awarded a provincial academic scholarship.

Delimitations

In completing the current study, the research was delimited in terms of the number of examinations analysed. Only a subset of the academic examinations administered in British Columbia was used in the current study. Latin 12, German 12, Spanish 12, Japanese 12, and all examinations written during the November, April, or August sittings were excluded from the study because fewer than 1000 students, generally considered the minimum number necessary to obtain stable parameter and ability estimations using the IRT models considered in the present study, wrote these examinations. The English 12 examination was not included in the study since it did not have a separate scholarship component during the two years being analysed and thus changes had not been made to the format of the English examination. French 12 was not included in the current study for two reasons. First, during the two years being analysed, the French 12 exam was also written by French Immersion students who were not enrolled in French 12. These students did not have a SBM for French 12 and could not receive credit for the course but they could use the results from the examination to obtain an examination scholarship score. This prevented the inclusion of SBM as auxiliary information for the French 12 examinations. Second, the June 1995 French 12 scholarship examination was not marked because it was made available to some students before the examination date, thereby compromising the results. June 1995 French 12 examination scholarship scores were based solely on the provincial examination.

History 12, Français Langue 12, and English Literature 12 also were not analysed. While these examinations did have a sufficient number of students for analysis, the major ER item in each examination was scored holistically using one or more scales and at least two markers. The holistic scales used five point scales that were then multiplied to increase the score value of the items, usually to 20 or 30. Unfortunately, the Ministry of Education only recorded the final weighted mark and it was not possible to determine the actual holistic scores given on these items. Since it was not possible to determine the actual holistic scores students received, these items could not be adequately analysed.

Due to the upper limit of 15 score categories in PARSCALE 3.1 (Muraki & Bock, 1997), it was necessary to reduce the number of score categories on specific polytomously-scored items. For example, in Biology there were examination items that were scored out of 10. However, it was possible for a student to receive scores separated by half-point intervals, for example, 3.5. If more than 15 separate score categories were found for a specific item, or the frequency of students receiving a score point was less than 0.2% for a specific item, these scores were rounded to the next score point.

CHAPTER 2

REVIEW OF THE LITERATURE

Chapter 2 is organized in seven sections. Section 1, The British Columbia Model, is focused on the unique aspects of the provincial examination system in British Columbia. The provincial examination program is described in greater detail, including the purpose, construction, administration, scoring, reporting of both the provincial and scholarship examinations, and the potential problems associated with the current procedure. The psychometric models considered in the present study are described in the second section. A comparison of classical and item response theory (IRT) is then presented followed by an examination of the use of IRT focusing on the issues associated with the use of unidimensional IRT models. The chapter concludes with three sections in which the research that has examined the value of extended-response (ER) items, the combined use of multiple-choice (MC) and ER items, and the use of subtest weighting and auxiliary information is reviewed.

The British Columbia Model

The current British Columbia provincial examination program has been in operation since the 1983/84 school year. While the purpose of the provincial examinations expanded in the 1996/97 school year to include the awarding of scholarships, the format and construction of the examinations did not change. As previously described, the expanded purpose of the provincial examinations is that the scholarship scores are now based solely on the provincial examinations because the separate one-hour optional provincial scholarship examinations were removed to save costs.

Purpose of the Provincial Examinations

As part of the High-School graduation requirements, students in British Columbia must complete a selection of required and elective courses at the grade 11 and 12 levels. Student grades for grade 11 courses are based entirely on the marks given by the classroom teacher. Both non-academic and locally developed grade 12 courses are graded in the same way. On the other hand, provincially developed grade 12 academic courses are graded differently. All grade 12 students in British Columbia must write a three-hour provincial examination in English 12 or Communication 12 (beginning with the 1999/2000 school year these two examinations were reduced to two hours in length) as well as a two-hour provincial examination in each of the academic grade 12 courses in which they are enrolled. In these courses, 60% of a student's grade is based on the SBM given by the classroom teacher and 40% on the student's provincial examination score.

However, the scholarship recipients are also determined using the provincial examinations. Before the 1996/97 school year, the scholarship scores for the academic courses, with the exception of English 12, were based on a summed score of the compulsory provincial examinations and the optional one-hour scholarship examinations. In the case of English 12, the scholarship scores were solely based on the three-hour provincial examination (Communications 12 is not part of the scholarship program). Beginning with the 1996/97 school year the scholarship examinations were removed and now all scholarship scores are based solely on the provincial examination results in all of the provincially examinable courses (except Communications 12).

Purpose of the Scholarship Examinations

Prior to the 1996/97 school year, optional scholarship examinations were in place for all of the provincially developed academic courses except English. The purpose of these optional examinations was to yield data which, when combined with performance on the provincial examinations, would determine provincial scholarship recipients. These examinations had no bearing on student course or examination grades. The scholarship examinations served two functions. First, they were a screening device since students who wanted a provincial academic scholarship had to write the course specific scholarship examination as well as the provincial examination in order to receive an examination scholarship score. Second, since individual student scores on both the scholarship and provincial examinations were combined to obtain individual scholarship scores, the scholarship examinations provided a greater range and discrimination amongst students in the raw scholarship scores than the provincial examinations alone would provide.

Construction of the Provincial Examinations

Despite the policy changes that occurred at the beginning of the 1996/97 school year, the provincial examinations did not change in content, construction, or format. Provincial academic courses all have a detailed curriculum guide containing the prescribed learning outcomes. The Ministry of Education uses these guides to annually develop and distribute a Table of Specifications detailing the examination content and format. Representative questions for each examination are provided at the same time to further clarify what is to be expected. These materials are provided to subject area teachers at the beginning of each school year.

The examinations assess a broad range of student achievement with respect to the curriculum of each course. Typically, the majority of the items used in the examinations are those that assess the concepts that are considered essential to that specific course. Most of the questions are moderately easy to moderately difficult ($0.45 \le p \le 0.75$). However, the examinations also contain some items that are conceptually and cognitively less difficult ($p \ge 0.75$) and a few that are more difficult ($p \le 0.45$). This distribution is used to increase the spread of scores for the purposes of student grading.

Every year, a number of teachers are contracted to develop a selection of both MC and ER items for the subject area in which they teach (e.g., Biology 12, Physics 12) and then construct examination forms using these items for potential use as provincial examinations. Typically, these teachers have extensive experience teaching and marking previous provincial examinations in the subject area. The teachers work independently to develop test items for a specific curriculum strand or curriculum focus that would elicit cognitive skills at one of three cognitive levels (Knowledge, Understanding, or Higher). Each test item must be explicitly linked to at least one learning outcome within that part of the curriculum. Item development guidelines are provided to each item writer. MC items must contain four alternatives, only one of which is correct or, in the case of best answer items, one that is considered the best. ER items must be clearly written using the standard rules of English such that the question provides the necessary information for each student to answer the question. The item writer also produces the solution to the question and the suggested value of the question.

Once the items have been developed, the item writers work together to construct the provincial examinations. Depending on the subject, between two and six forms are developed such that there is a form for each sitting of the examination for the upcoming year. Each form is only used once and there is one extra form in case of a contingency (e.g., an examination is compromised). The item writing team members begin by reviewing the items for correctness and appropriateness, modifying or removing items as necessary. The items are then placed into one of the forms in such a way that the set of items within each form fits the Table of Specifications and the forms are considered approximately equal in terms of difficulty. Such decisions are based on the professional opinion of the item developers because little or no field-testing is completed. Unless there is a change in course curriculum or focus, the same Table of Specifications is used each year. Consequently, the content and format of the examinations remain "essentially" constant across time. This includes the number of items and marks allocated to each curricular strand, the expected cognitive levels of thinking required, and the order of the placement of the curricular strands within each examination.

Table 1 contains the allocation of marks for MC and ER items and a brief description of the format of the ER section for each examination included in the present study. As shown in column 2, the scoring for each of the six subjects varies in the number of marks allocated to either the MC or ER sections as well as in the total marks allocated for each examination. For example, 52 marks are allocated to the MC section and 48 marks to the ER section for a total of 100 marks in the Biology examinations. In contrast, 48 marks are allocated to the MC section and 32 marks to the ER section for a total of 80 marks in the Chemistry examinations. Column 3 provides a brief description of the format of the examinations. As shown in column 3, the number of ER items varies from 6 to 14 and the value of the ER items varies from 2 to 10 across the examinations. All of the ER items are compulsory except in Biology and Physics. In these two subjects, the ER sections consist of a set of compulsory items and a set of optional items from which the student must answer two in the case of Biology and one in the case of Physics. Physics is also unique because the MC items are worth two marks each instead of one mark as in the other examinations.

Table 1

Value, Format, and Description of the Provincial Examinations used in the Present Study

Course	Scoring	Format
Biology 12		5 to 7 compulsory ER items worth 28 marks. Item values vary from 2 to 6. Two other 10-mark items must also be selected from choice of 7 specialized topics. Students are expected to define key terms, describe functions, and solve related conceptual problems.
MC	52	
ER	48	
Total	100	
Chemistry 12		11 to 13 compulsory ER items with item values of 2 to 6. Students are expected to define key concepts and solve mathematical or conceptual problems.
MC	48	
ER	32	
Total	80	
Geography		12 to 14 compulsory ER items with item values of 2 to 6. Students are expected to define key concepts and solve conceptual problems.
MC	40	
ER	60	
Total	100	
Geology		12 to 14 compulsory ER items with item values of 2 to 8. Students are expected to define key concepts and solve conceptual problems.
MC	60	
ER	40	
Total	100	
Math		6 or 7 compulsory ER items with item values of 2 to 5. Students are expected to solve mathematical or conceptual problems.
MC	50	
ER	20	
Total	70	
Physics		30 MC items valued at two marks each. 7 to 8 compulsory ER items worth 48 marks. Item values vary from 2 to 5. One other 12-mark item with 3 components must be selected from 3 specialized topics. Students are expected to define key concepts and solve mathematical or conceptual problems.
MC	60	
ER	60	
Total	120	

Construction of the Scholarship Examinations

The same developmental process employed to construct the provincial examinations was used to develop the scholarship examinations when they were in place. However, only ER items were used and the items used in the scholarship examinations were considered conceptually and cognitively more complex than those used in the provincial examinations. Thus, the scholarship examinations focused on complex concepts rather than the core concepts of each subject's curriculum. Most of the questions were quite difficult with few, if any, easier questions.

Administration of the Provincial Examinations

At the beginning of the school year, the Ministry of Education develops an examination calendar that gives the date and time for the writing of all provincial examinations. Currently, all examinations are completed within a specified time period at the end of each quarter, semester, and school year. This examination period is 2 days in the case of the November and April sittings, 5 days for the January sitting, and 7 days for the June sitting. The difference in time allotment is due to the number of examinations to be administered. For each subject area, the administration of the provincial examination occurs at the specified time and date at a neutral location in those High Schools in which the course is offered.

School officials are responsible for informing the Ministry of Education which provincially examinable courses are being offered at the school and the number of students who will likely be writing each examination. This information is then used to distribute the required number of examinations to each school along with student identification tags containing student names and provincial student ID numbers. Teachers at the school invigilate each examination, although not the teachers who teach the course. During each examination, the invigilators place the identification tags on the corresponding students' examinations and MC response forms. With the exception of English and Communications 12, students have two hours to complete each examination.

Administration of the Scholarship Examinations

The same administration procedure was used for the scholarship examinations. During the school year, students interested in writing scholarship examinations informed the designated school representative, usually a school counselor, of their intention. This information was forwarded to the Ministry of Education and then used to distribute examination booklets and student identification tags. Each scholarship examination began 15 minutes after the completion of the corresponding two-hour provincial examination and was one hour in length.

Scoring of the Provincial Examinations

The MC items are scored using optical scanners. As indicated previously, for the examinations considered in the present study, one point is awarded for each correct response and zero points awarded for incorrect responses except for Physics. For Physics, two points are awarded for each correct response and zero points for each incorrect response.

Subject area teachers, contracted by the Ministry of Education, score the ER items. These teachers work as marking committees under the supervision of a committee chairperson who is a teacher with several years of experience teaching the course and marking previous examinations. The size of the marking committee varies depending on the number of examinations to be scored. Committee members are selected in order to have teachers with varying amounts of marking experience, from novice to very experienced, with representation from different geographical locations within the province. For each course, teachers meet at a central location the week following the provincial examination period.

Both analytic and holistic scales are used to mark the ER items; however, only analytic scoring is used on the examinations considered in the present study. In these examinations, the ER items are restricted-response essays, definitions, or problems. Student responses are compared to a key and graded based on the level of completeness as compared to the key. Although it is not usually necessary for the response to be identical to the key, a student's response must be such that it can be compared to the key in order to receive full credit, partial credit, or no credit. Students receive partial credit depending on the completeness of the response. This partial credit can be in increments of one-half or one score point.

Some formal training is provided to those teachers using holistic marking scales (e.g., English 12). Experienced and novice markers are also mixed in order to provide further informal training. In examinations having analytic scales, the key specifies the allocation of partial marks. A subset of markers is assigned to mark one or two ER items. These markers work together to fine tune the key and maintain consistency throughout the marking session. Thus, different markers mark different items and each item is marked by a single marker. In the case of holistically scored items, two markers mark each item and on rare occasions (i.e., if the two markers produce very different scores for the same response), a third marker also scores the item.

Besides marking the ER portion of the examinations, the marking committees also complete two other tasks. First, they review the MC items in terms of structure and item statistics. Items with no correct responses or more than one correct response are deleted and the examination total is reduced accordingly. Items with poor item statistics are reviewed and, if a logical explanation for the poor fit based on course content can be ascertained, the marking committee suggests to the Ministry of Education that the item be removed. The Ministry of Education then decides either to keep or remove the item. ER items are rarely if ever removed. However, the scoring keys are modified to compensate for question ambiguity or unexpected but conceptually defensible alternative responses.

Second, the members of each committee work independently, using their professional judgement regarding the difficulty of the examination and the proportion of students receiving each letter grade based on the raw examination scores, to suggest the examination scores that correspond to each of the letter grade cut points. In British Columbia, a score of 86% or higher is considered an A, 73 to 85% a B, 67-72% a C+, 60 to 66% a C, 50 to 59% a P, and below 50% an F. The mean of the marker ratings, after a second iteration, is used to make slight adjustments to the raw examination scores, thereby adjusting examination percentages and the proportion of students obtaining each letter grade on the examination. For example, a raw score of 84 out of 100 on a provincial examination may be adjusted to be reported as 86%. Likewise a score of 73 out of 100 may be reported as 72%. These adjustments are made to compensate for variations in the overall difficulty across the different forms of the examinations.

Scoring of the Scholarship Examinations

The same marking committees that scored the provincial examinations completed the scoring of the scholarship examinations. After the provincial examinations were completely scored, the committee then scored the scholarship examinations using the same procedure as used for the ER items on the provincial examinations.

Reporting of the Provincial Examination Results

Item level student scores are entered using optical scanners for the MC items and data entry personnel for the ER items. Computer programs are used to calculate the total raw score and the reported examination score, which is a rounded percentage score based on the adjusted raw score (based on the adjustments recommended by the scoring committee). Reported examination scores are combined with the school-based percentages, weighted 40/60, respectively, to determine course grades and percentages. No adjustments are made to account for possible differences in variability which, when left unaccounted for, influence the weights used. Approximately six weeks after writing provincial examinations, students are informed of their provincial examination and course results.

Reporting of the Scholarship Results

Since only ER items were used in the scholarship examinations, data entry personnel entered item level scores. Computer programs were used to calculate the total raw score. Once all of the provincial and scholarship examinations for a course were completely scored, each student's raw score on both the provincial and scholarship component were simply summed together to obtain a raw scholarship score and rounded to the nearest whole number. The RANKIT procedure ((r-1/2)/w, where r is the rank and w is the sum of the case weights, was then used to change the scholarship score distribution such that it resembled a normal distribution (Chambers, Cleveland, Kleiner, & Tukey, 1983). These ranked scores were then rescaled such that the mean and standard deviation of the set of scholarship scores for each course were 500 and 100, respectively. Further, the minimum and maximum scores were set to 200 and 800 and those scores below 200 or above 800 were rounded to 200 or 800, respectively. If the maximum scholarship score for a specific course was below 800, all of the scores above 675 were adjusted using what is called the Kozlow corollary, a correction formula developed and used by the British Columbia Ministry of Education to ensure that the highest scholarship score for each course was 800, three standard deviations from the mean. The formula for the Kozlow corollary (Glenn Church, personal communication, February 22, 2000) used to determine the adjusted scaled scholarship score ($SMS_{ij}*$) for student *i* on examination *j* is:

$$SMS_{ij}* = 675 + 125\left(\frac{(SMS_{ij} - 675)}{(SMS_{\max j} - 675)}\right), \qquad (1)$$

where SMS_{ij} is the scaled scholarship score for student *i* for examination *j*, where $SMS_{ij} \geq 675$, and

$SMS_{max\,j}$ is the maximum calculated scholarship score for the group of students writing examination *j*.

The current procedure uses only the provincial examination scores. Only those students with provincial examination scores of 70% or more are used to calculate the scholarship scores. However, based on the scores of this subset of students, the RANKIT procedure, scale transformation, and, if necessary, the Kozlow correction formula are applied as before. Thus, the examination scholarship scores range from 200 to 800 and have a mean of 500 and standard deviation of 100.

Once the scholarship scores are calculated and recorded, they are sent to students and the school administration at the same time that the provincial examination results are released. The examination and total scholarship scores are provided to all students who have at least one scholarship score. The Ministry also maintains a record of the total scholarship score obtained by each student based on their three highest scholarship scores above 475. If a student's total scholarship score is at least 1700, the student is informed, along with the school administration, that the student is entitled to a provincial academic scholarship.

Potential Problems with the Current Procedure

The removal of the scholarship examinations has resulted in the removal of a large portion of the more challenging ER items, thus decreasing the length and the difficulty of the examinations used to determine the scholarship scores. Consequently, it seems likely there will be differences in the scholarship scores and decisions between the original and current procedures. Figure 1 compares the scholarship scores that were calculated using the original and current procedures for the 1994/95 June Chemistry examination. When the original procedure was in place, a range of scholarship scores was associated with each provincial examination score because the scholarship examination provided further separation amongst students having the same provincial examination score. In contrast, with the current procedure, a single scholarship score is associated with each provincial examination score. At each provincial examination score point, the scholarship score from the current procedure is generally near the midpoint (except at the low end of the distribution) of the scholarship score range from the original procedure. The distribution of the scholarship scores for each provincial examination score is nonlinear because of the ranking procedure and correction formula. The correlation between the results of the two procedures is 0.94. However, for a given provincial examination score point, the range of the scholarship scores based on the original procedure was approximately 200 points, especially at the higher end of the score scale. Thus, the use of the current procedure could translate into relatively large differences in both the examination and total scholarship scores as compared to the original procedure.

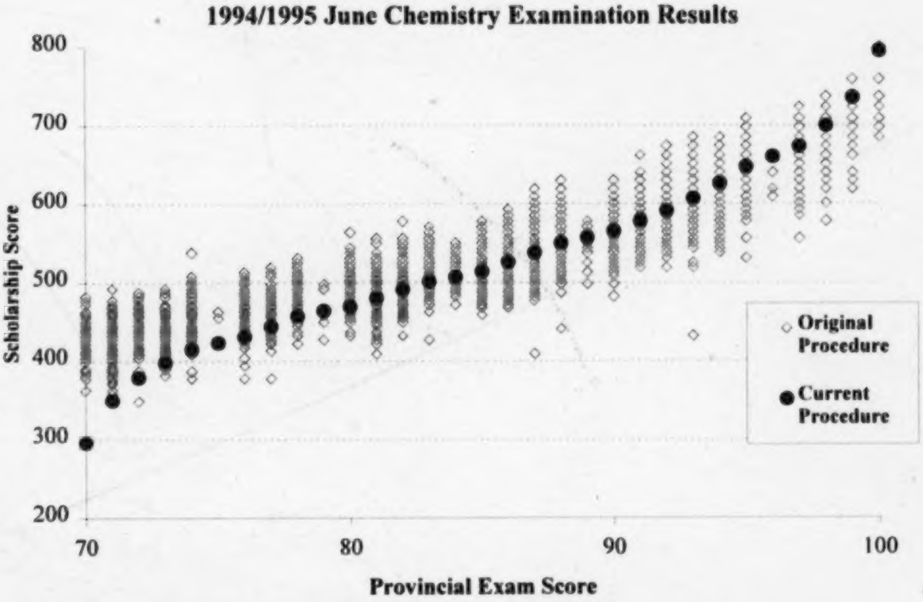


Figure 1. Comparison of the Scholarship Scores Calculated using the Original and the Current Procedures

Other potential problems also exist. Students who generally do better on ER items likely will be more negatively affected than those who do better on MC items. Research indicates that males do better on MC items while females do better on ER items (e.g., Bolger & Kellaghan, 1990, Garner & Engelhard, 1999; Henderson, 1999). Based on the examination results of 15 year old males and females on Mathematics and English examinations, Bolger and Kellaghan (1990) found gender–by–item type interactions regardless of subject area, with boys having a relative advantage on MC items and females having a relative advantage on ER items. More recently, studies of differential item functioning for grade 11 and 12 students in both Canada and the USA have found that for students of equal ability, MC items will more likely favour males while ER items will more likely favour females (Garner & Engelhard, 1999; Henderson, 1999).

Nevertheless, some contradictory evidence also exists. In the study completed by Bridgeman (1992), in which Graduate Record Examination MC items were changed into dichotomously-scored ER items, no gender–by–item format interactions were observed. DeMars (1998) and O'Neil and Brown (1998) obtained similar results. However, of interest for the present study is a secondary finding by DeMars (1998). While DeMars did not find any overall differences, she did find that for the highest ability students (top 5%), the gender differences described above were detected, a finding that coincided with previous research conducted by Bridgeman (1989) and Schmitt, Mazzeo, and Bleistein (1991). To the extent that the scholarship program in British Columbia is for higher achieving students, these findings further suggest that the current procedure may have a previously unanticipated differential impact on higher achieving males and females.

Psychometric Models

Classical Test Score Theory

The provincial examination program in British Columbia uses the total test mark obtained by a student as an estimate of achievement with respect to the construct (trait) being measured. While the origins of testing can be traced back to China in 2200 B. C. (Lien, 1967), it was not until 1904 that Spearman proposed that the observed total test score was a composite score consisting of a true score and an error score due to what Spearman called error of measurement. According to this model, X_{if}, the observed score for examinee i on examination f, can be expressed as:

$$X_{if} = \tau_i + \varepsilon_{if}, \qquad (2)$$

where τ_i is the true score for examinee i on the variable of interest, and

ε_{if} is the error of measurement for examinee i on examination f (p. 107, Crocker & Algina, 1986).

Within the classical framework, the observed score, X_{if}, is an unbiased estimate of τ_i. Thus, the observed score is commonly used in testing programs like that used in British Columbia. The observed scores are reported and used not only to estimate the achievement of each student, but also to rank the students. Although different methods can be used to obtain the observed score, the British Columbia provincial examination program uses the simple sum of item scores to obtain this score. Hence, each raw score point contributes equally to the total observed score for each student. The one variation used in determining the British Columbia scholarship scores is that the summed examination test scores are ranked and normalized using the RANKIT procedure before they are transformed into examination scholarship scores.

Item Response Theory (IRT)

Although research into the use of IRT as an alternative to classical test score theory has mostly occurred over the last three decades, the foundations for IRT can be traced to the early work completed by Binet and Simon in 1916 and Richardson in 1936 (Hambleton & Swaminathan, 1985). However, Lord (1952) is generally credited with providing the general framework for IRT. The fundamental difference between classical test score theory (CTST) and IRT is that CTST operates at the examination level while IRT operates at the item level. Thus, using IRT, an examinee's "observed score" is calculated using the item characteristics and the response pattern of the examinee. Within the framework of IRT the term ability (θ) is used instead of observed score to signify that it "is a label used to designate the trait or characteristic that a test measures" (Hambleton & Swaminathan, 1985, p. 54). Thus, the ability score is an estimate of an examinee's latent trait (domain score). Admittedly, the term is a convenience that requires studies of validity in order to link the estimates and the construct of interest (Hambleton & Swaminathan, 1985). As with the true score model that forms the basis of classical test score theory, IRT models generate the most probable solution to an indeterminate problem, in this case the parameters defining the item response functions and the estimates of examinee ability.

IRT is represented by a class of mathematical probability models (functions) described below that use the examinee item response vectors to estimate the item level parameter(s) for each item to best fit the distribution of the examinee responses and scores. The item parameters are then used to produce θ estimates for the examinees, through a series of iterations that best fit the item level response data.

Dichotomous Item Response Models

The first IRT models were developed for the analysis of dichotomously-scored items (e.g., multiple-choice items). The three most commonly used dichotomous IRT models (one-, two-, and three-parameter models) are distinguished by the number of item parameters to be estimated. Although only the two- and three-parameter dichotomous models will be considered in the present study, the one-parameter model is provided first as a foundation upon which the other models have expanded.

One-parameter dichotomous item response model. The one-parameter dichotomous item response model is defined by the logistic probability function in which $P_j(\theta_i)$, the probability that examinee *i* with an ability of θ will correctly answer item *j*, is expressed by:

$$P_j(\theta_i) = \frac{1}{1 + \exp^{-1.7(\theta_i - b_j)}}, \tag{3}$$

where θ_i is the ability of the examinee *i*,

b_j is the θ value at the inflection point corresponding to the difficulty of the item and is at the point on the θ scale at which the probability of correctly answering item *j* is 0.50, and

1.7 (often labeled D) is the scaling factor that transforms the logistic model to be on the same metric as the normal ogive model albeit using less computational complexity (Hambleton & Swaminathan, 1985).

With the one-parameter model, higher values of *b* represent items that are more difficult to answer.

Two-parameter dichotomous item response model. The two-parameter dichotomous item response model is defined by the logistic probability function in which $P_j(\theta_i)$ is expressed by:

$$P_j(\theta_i) = \frac{1}{1 + \exp^{-1.7a_j(\theta_i - b_j)}}, \tag{4}$$

where a_j is the slope of the function at the inflection point and is commonly called the discrimination parameter, and

θ_i and b_j are defined as before.

Considered a constant across items in the one-parameter model, the presence of differing a-parameters across items allows for probability functions with different slopes. Larger a-parameter values indicate items with higher discrimination among examinees of differing ability near the inflection point.

Three-parameter dichotomous item response model. The three-parameter dichotomous item response model is defined by the logistic probability function in which $P_j(\theta_i)$ is expressed by:

$$P_j(\theta_i) = c_j \frac{(1 - c_j)}{1 + \exp^{-1.7a_j(\theta_i - b_j)}}, \tag{5}$$

where c_j shifts the probability functions vertically, thus creating a function with a lower boundary that is greater than zero, as is assumed with the previous two models,

b_j, while still being the θ value at the inflection point, occurs at a probability of $0.50 + 1/2c_j$, and

a_j and θ_i are defined as before.

Given that a positive lower boundary represents a non-zero probability for examinees at all ability levels to correctly answer the item, the c-parameter is commonly called the pseudo-guessing parameter since higher *c* values indicate the item is more likely to be answered correctly by examinees regardless of their ability.

Item response curve (IRC). IRCs are graphical representations of the probability functions described above. Figure 2 displays two typical s-shaped IRCs for the three-parameter model illustrating the effects of different item parameters on the shape and location of the curve. The x-axis is the θ continuum transformed onto a scale with a mean of zero and standard deviation of one. The y-axis is the probability of correctly answering the item. As shown in Figure 2, an item with a higher *b*-parameter value (item 2) has an inflection point further to the right representing a more difficult item. An item with a larger *a*-parameter (item 2) has a steeper curve and better differentiates among examinees with θ values near the inflection point. Finally, an item with a higher *c*-parameter (item 1) has a higher lower asymptote indicating a higher likelihood of guessing on the item.

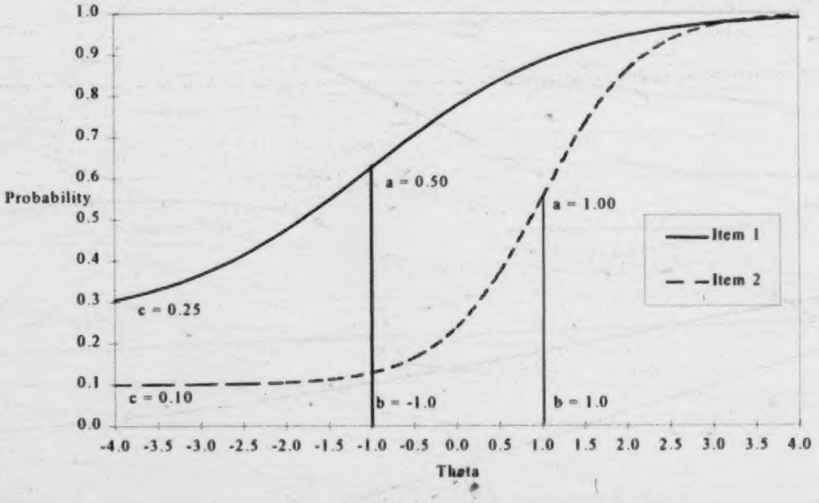


Figure 2. Item Response Curves for Two Dichotomously-Scored Items

Given that a positive lower boundary represents a non-zero probability for examinees at all ability levels to correctly answer the item, the c-parameter is commonly called the pseudo-guessing parameter since higher *c* values indicate the item is more likely to be answered correctly by examinees regardless of their ability.

Item response curve (IRC). IRCs are graphical representations of the probability functions described above. Figure 2 displays two typical s-shaped IRCs for the three-parameter model illustrating the effects of different item parameters on the shape and location of the curve. The x-axis is the θ continuum transformed onto a scale with a mean of zero and standard deviation of one. The y-axis is the probability of correctly answering the item. As shown in Figure 2, an item with a higher *b*-parameter value (item 2) has an inflection point further to the right representing a more difficult item. An item with a larger *a*-parameter (item 2) has a steeper curve and better differentiates among examinees with θ values near the inflection point. Finally, an item with a higher *c*-parameter (item 1) has a higher lower asymptote indicating a higher likelihood of guessing on the item.

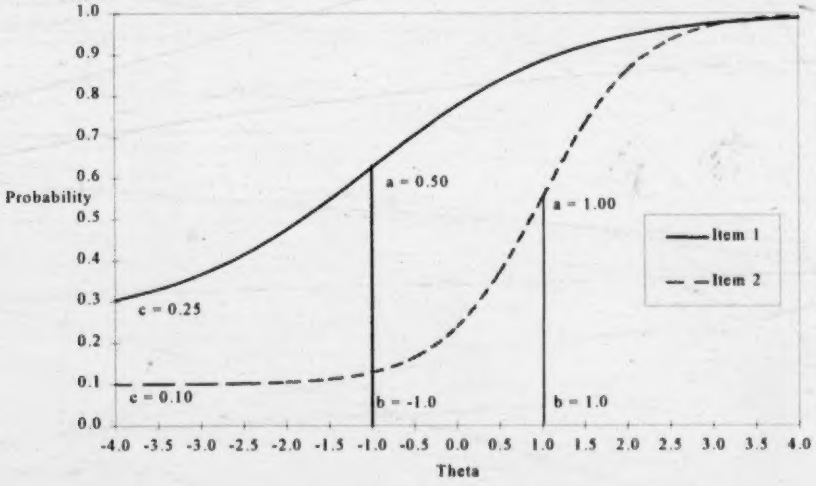


Figure 2. Item Response Curves for Two Dichotomously-Scored Items

Polytomous Item Response Models

While the three previous models are useful for the analysis of dichotomously-scored items having two score categories, they can not be used when items contain more than two score categories. In the case of polytomously-scored items, different but related models have evolved. Polytomous item response models are a generalized form of the dichotomous models because they are based on mathematical functions that estimate item parameters and ability using items with score scales having two or more score points.

The generalized partial credit model (GPCM). The GPCM is a polytomous model developed by Muraki (1992). The model is an extension of Masters' one-parameter partial credit model (1982) and a formulation of the generalization of Masters' model as first proposed by Thissen and Steinberg (1986). Since the GPCM allows for differing a-parameters across items, it is considered a two-parameter model for polytomously-scored items. In fact, the GPCM reduces to the two-parameter dichotomous model when dichotomously-scored items are used. However, when used with polytomously-scored items, the a- and b-parameters function somewhat differently. The a-parameter represents discrimination and the b-parameter represents difficulty in a dichotomously-scored item. In the GPCM, both the a-parameter and the set of threshold parameters determine discrimination. Since there are k-1 response curves in the GPCM, where k is the number of response categories, the b-parameter, called the item step parameter, is subscripted b_{jv} and represents the location on the θ scale where the probability of two adjacent score categories intersect (Muraki, 1992). The probability function $P_{jk}(\theta_i)$, the probability that examinee i with a given θ value will achieve category k on item j, for the GPCM is:

$$P_{jk} = (\theta) = \frac{\sum_{v=1}^{k} a_j(\theta - b_{jv})}{\sum_{c=1}^{m_j} \exp\left\{\sum_{v=1}^{c} a_j(\theta - b_{jv})\right\}}, \qquad (6)$$

where m_j is the number of possible score categories,

v is the score category being analysed,

c is the score categories 1, 2, ..., m_j,

b_{jv} is the item category parameter, the ability at which a category score of k or k-1 is equally likely, and

a_j is the discrimination (slope) parameter (Muraki, 1992).

By convention b_{j0} is arbitrarily set to 0.0 (Muraki, 1992) since this term is canceled out of the numerator and the denominator. The GPCM is a "divide by total" model because the denominator in the function represents the total amount of information provided by a specific item (Thissen & Steinberg, 1986). Furthermore, this division normalizes the category probabilities so that the maximum probability is 1.0.

Since its introduction, the GPCM has been found to produce good approximations of the actual parameter and ability estimates under simulated conditions and is being increasingly used in the measurement of polytomously-scored ER items (e.g., Muraki, 1992, 1993; Donoghue, 1994; Carlson, 1996; Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996). In comparison to the one-parameter partial credit model, the GPCM has been shown to produce superior estimates in both simulated and actual testing conditions (Fitzpatrick et al., 1996). Previous research has also shown that the GPCM is comparable to the two-parameter graded response model that was introduced in 1972 by Samejima (Maydeau-Olivares, Drasgow, & Mead, 1994; Klinger & Boughton, 1999). Thus, the

GPCM is considered a viable unidimensional IRT model for parameter and ability estimation for examinations containing polytomously-scored items.

Item category response curve (ICRC). An ICRC is a graphical representation of a polytomous probability function. For each polytomously-scored item, k-1 ICRCs are computed. Figure 3 illustrates a set of ICRCs for an item with four possible scores. For examinees of increasing ability, the probability of receiving a score in the lowest category decreases while the probability of receiving a score in the highest category increases. The width of each ICRC is dependent on the *a*-parameter. Items with larger *a*-parameters will have steeper narrower ICRCs with less overlap between score points than items having smaller *a*-parameters. Less overlap represents better discrimination between score categories. The intersection of adjacent ICRCs is the ability at which two adjacent score categories are equally probable. For a given item, there is only one *a*-parameter value but the number of b_{jv} values is dependent on the number of score categories. Finally, at any level of θ, the sum of the probabilities on the set of ICRCs is one.

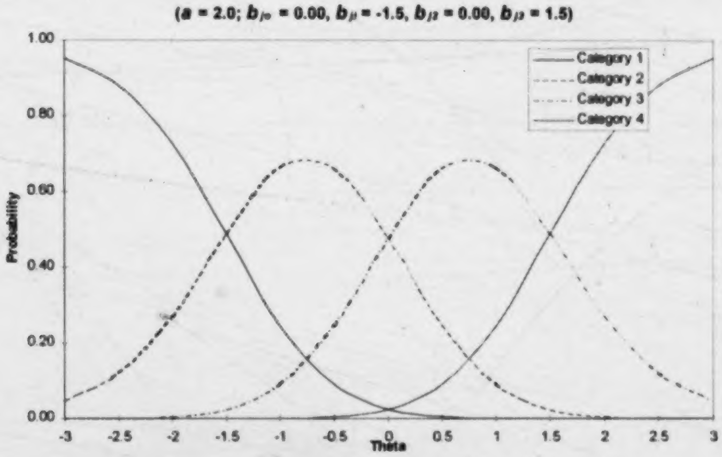


Figure 3. Item Category Response Curves for a Four-category Item

As with any model, the ability of the GPCM to produce reliable results is dependent on the distribution of responses across the alternative scores. If a given score category does not contain many responses, the ICRC for that category will be much lower than that of the other categories (Muraki, 1993). Muraki further determined that low response categories often reduced the precision of the item parameter and ability estimates. Using National Assessment of Educational Progress (NAEP) writing items, he demonstrated that score categories with a low level of responses could be combined with adjacent categories with no loss in item information. In many cases, the combining of these categories actually increased the amount of information an item could provide towards ability estimation. Based on these results, the combining of low response categories with adjacent categories should not affect the precision of the ability estimates. Thus, an important step in the use of the GPCM is to analyse the data for low response categories and then combine these low response categories with adjacent categories.

A Comparison of Classical Test Score Theory and Item Response Theory

Although defined differently and not synonymously with latent trait (ability), the true score, as estimated by classical test score theory, is theoretically related to the latent trait, as estimated by the ability (θ) estimates in IRT. The relationship between τ and θ is nonlinear and the distributions have different shapes (Lord, 1953, 1980). Since the theoretical values derived from the different approaches are related, the merits of each theory should be based on the quality of the estimates produced by the models used to operationalize the theory. The estimates using IRT should be different and somewhat superior to those derived from the classical theory since both item and examinee response vectors, in the two- and three-parameter models, are used to determine the ability estimates for each examinee (Thissen, Pommerich, Billeaud, & Williams, 1995).

Previous research has not been able to consistently support these hypotheses. While Birnbaum (1968) first demonstrated that the scores based on response patterns differed from those based on the summed scores, other research has shown the differences to be small. Fan (1998) illustrated this using the grade 11 Texas Assessment of Academic Skills (TAAS) examination to compare both the classical and IRT frameworks. Based on either a two- or three-parameter model, Fan found the correlations between the IRT models and the total test score to be at least 0.96 in all cases and concluded that the same or very similar conclusions would be drawn regardless of the method used. Anderson (1999) found similar correlations when he compared the three-parameter IRT model and the total test score for the dichotomously-scored items on the January 1996 British Columbia Mathematics provincial examination. However, this previous research has been limited in that the comparisons have only been completed for examinations having dichotomously-scored items. Further, even small differences can be important if they are shown to be consistently superior.

One of the difficulties in comparing the classical and the IRT frameworks is to develop a standard of comparison to compare the results. Within the IRT framework, the use of simulated research has been used to measure the utility and superiority of different models (e.g., Reise & Yu, 1990). However, it is difficult to compare both the classical and IRT models using simulated conditions. One approach for comparing the superiority of these models using actual examination data is the use of a shortened version of the full examination (Bock, Thissen, & Zimowski, 1997; Anderson, 1999; Folske, Gessaroli, &

Swanson, 1999). The comparisons of the examinee score estimates derived from the shortened examination to those of the full examination provide a measure of the accuracy of the procedure employed to obtain the estimates. Bock, Thissen, and Zimowski (1997) used 20 item subtests from a 100 item spelling examination to compare the two approaches. Using examinees' scores on the 100 item examination as a standard for comparison, the examinees' ability estimates from the 20 item subtests as determined by the two-parameter IRT model were closer to the standard of comparison than the percent correct score on these subtests.

In contrast, Anderson (1999) did not find a difference in superiority. He created two subtests comprising of either the odd or even numbered items from the 50 multiple-choice items of the January 1996 British Columbia Mathematics provincial examination. The domain scores as estimated using the total scores expressed as proportions or the θ estimates on these subtests were compared to the actual provincial examination scores. His conclusion, while admittedly exploratory, was that the scores derived from the two procedures were almost identical in terms of the means, standard deviations, correlations, and classification decisions. However, closer analysis of the results reported by Anderson does illustrate some small but notable differences. For example, the mean domain scores for each method were the same but the standard deviation for those scores derived from the three-parameter IRT model was less (see Anderson, 1999, Table 1, p. 348). Further, the root mean square error values were marginally larger for the three-parameter model (see Anderson, 1999, Table 3, p. 349). Finally, small differences did exist in the assignment of letter grades to students between the approaches with the three-parameter model providing marginally superior assignment (see Anderson, 1999, Table 5, p. 350).

Nonetheless, this previous research has been limited. First, the studies only examined one or two examinations. Second, and more importantly, the results were only based on the analysis of dichotomously-scored items. Although Anderson (1999) used the reported examination score, which is based on both dichotomously- and polytomously-scored items, as the standard for comparison and Bock et al. (1997) briefly discussed the implications for polytomously-scored items, none of their work extended to the comparison of classical test score and polytomous IRT models. Samejima (1996) has completed some exploratory work in this respect and has shown that the use of response patterns produces superior results than the use total test score for examinations consisting of polytomously-scored items. However, research comparing the models with actual examination data has not been completed nor has there been any comparative work using examinations that consist of both dichotomously- and polytomously-scored items.

The Use of Item Response Theory

Unlike classical test score theory, which is based on a sum of the item scores to estimate the true score, the IRT estimation procedure is more computationally complex and the derived estimates vary depending on the choice of model (Anderson, 1999). Thus, the IRT model chosen must represent the data for which it is to be based. Previous research has focused on the assumptions that must be met in order to choose and justify the use of the various IRT models.

IRT Assumptions

The choice and justification of the IRT model to be used is predicated on the satisfaction of the underlying assumptions for that model. The major assumptions to be tested are unidimensionality, local item independence, nonspeededness, and lack of guessing. While it is unlikely that a set of data will ever fully meet the required assumptions, the degree of fit between the model and the data will affect the quality of the item parameter and θ estimates (Traub, 1983). Thus, it is important to examine the criteria for these assumptions and the effects that can be expected when the assumptions are not fully met.

Unidimensionality. Since the θ estimates provided by the IRT models are an estimate of each student's latent trait, it is important that the examination items are essentially measuring a single trait. Evidence of unidimensionality justifies the use of an unidimensional model and also provides evidence of local item independence and, to a lesser extent, nonspeededness and lack of guessing.

One issue is how to determine unidimensionality. Given the warning of Traub (1983) and the likelihood that the items included in an examination are actually measuring a series of closely related traits, essential unidimensionality has been proposed as a sufficient measure of unidimensionality (Stout, 1990). Under the assumption of essential unidimensionality, the data are dominated by one dimension with the other dimensions having a weak influence (Stout, 1987). Historically, factor analysis has been used to determine dimensionality of a given data set (Hambleton & Swaminathan, 1985). Using factor analysis, researchers have concluded that an examination is essentially measuring an unidimensional trait if there is a dominant first factor in the data set as indicated by a high ratio between the first and second eigenvalues as compared to the ratio between the other pairs of eigenvalues (Reckase, 1979; Gorsuch, 1983; Ndalichako, 1997). The ratios of the other pairs of eigenvalues should also be close to one. Graphically, a scree plot of these eigenvalues would only show a single dominant factor. Reckase (1979) also found that acceptable IRT calibrations occurred with dichotomous models if the first factor accounted for 20% of the total test variance and that good θ estimates could be obtained even if the first component accounted for only 10% of the variance.

While this criterion was based on examinations having dichotomously-scored items, Huynh and Ferrara (1994) have used it to consider examinations having polytomously-scored items to be essentially unidimensional. They also concluded that even if the cognitive processes required to answer ER items were inherently complex and multidimensional, the responses were so dominated by the first principal component, which in their study accounted for as much as a quarter of the variance, that the use of an unidimensional IRT model was appropriate (Huynh and Ferrara, 1994).

When the assumption of unidimensionality is not met, the estimation procedure within the unidimensional models can be expected to produce ability and item parameter estimates that reflect the different dimensions. Yen (1986) concluded that the use of unidimensional models with multidimensional data would produce ability and parameter estimates that were weighted means of the underlying parameters. These weights were proportional to the underlying item discriminations, such that the discrimination increased as the number of important traits increased. Previous research tends to support this conclusion. For example, based on simulated studies in which multidimensional dichotomous data were created for correlated traits (θ_1 and θ_2), the unidimensional ability estimates were found to approximate the mean of θ_1 and θ_2 (Way, Ansley, & Forsyth, 1988; Ackerman, 1989). Similar findings have been found with the use of polytomously-scored items. De Ayala (1994, 1995) used simulated data for which the latent space was determined by two correlated dimensions. As with the previous studies, he found the θ estimates derived using the unidimensional one-parameter partial credit model were a more accurate estimate of the mean of θ_1 and θ_2 rather than either θ_1 or θ_2 independently (De Ayala, 1995). Interestingly, while Yen (1986) admitted that no unidimensional method was strictly appropriate when items differed in dimensionality she reported that within the context of the dichotomously-scored Mathematics items she was analysing, "the (unidimensional) IRT model appeared to handle the multidimensional data in a reasonable manner" (p. 321).

Local item independence. The assumption of local item independence is based on the notion that a student's responses to different items in an examination are statistically independent, such that the performance on one item does not affect a student's performance on another item (Hambleton & Swaminathan, 1985). In this respect, local item independence is somewhat related to unidimensionality. If items are not statistically independent, a second factor is required to account for the performance of the examinees. Thus, if an examination has met the criteria for unidimensionality, it will also have met the criteria for local item independence. However, if an examination is found to be multidimensional, the assumption of local item independence needs to be assessed.

Examinations combining MC and ER have been shown to violate the assumption of local item independence (Thissen, Wainer, & Wang, 1994). The results of the IRT analysis can be problematic if local dependence exists. For example, the *a*-parameters will be overestimated causing an overestimation of the information function (Sireci, Thissen, & Wainer, 1991). One solution is to separate the locally dependent items into a testlet and estimate the parameters for the testlet separately from the rest of the examination (Yen, 1992). However, if the local dependence is small, as shown by the assessment of unidimensionality, or consists of a small number of items, it is likely that within an unidimensional IRT framework, these items could be combined with the other items with no adverse effects on the estimation process (Thissen et al., 1994).

Nonspeededness. Nonspeededness refers to the assumption that the majority of examinees (at least 85%) have sufficient time to attempt all of the questions and thus incorrect responses or omissions are due to lack of knowledge rather than lack of opportunity to attempt the questions. Nonspeededness can be assumed if the assumption of unidimensionality is met since a speeded test would have a second factor, speed, which would account for examinee performance. A second method to determine nonspeededness is to examine the item completion rate. To measure the completion rate, Hambleton and Swaminathan (1985) suggested a three part test: the percent of examinees completing the test, the percent of examinees completing 75% of the test, and the number of items completed by 80% of the examinees. One method to complete these tests is to examine the completion rate of the last items on an examination. For example, Ndalichako (1997) considered a 95% completion rate on the last three MC items as evidence that speed was not a factor. A further examination is to determine the proportion of examinees that did not omit any items over the examination. If this number is above 80%, nonspeededness can be assumed.

Lack of guessing. The presence of guessing without actually speaking to examinees is difficult to establish (Hambleton & Swaminathan, 1985). Nonetheless, psychometric procedures have been proposed that can suggest if guessing is likely or unlikely. First, a nonspeeded examination is generally considered to be less susceptible to guessing. Second, the performance of the lowest group of examinees on the most difficult MC items should be well below 25%. Ndalichako (1997) considered close to zero performance on the three hardest items by the lowest scoring examinees as evidence that guessing was not a factor.

If guessing is a factor, it can lower the fit of the data with the two-parameter model and prompt the use of the three-parameter model. However, this reduction in fit is usually only at the lower end of the ability distribution. Thus, the presence of guessing may only be a factor for those examinees in the lower ability ranges and would not generally affect the estimation procedure for higher ability examinees (Lord, 1980; Hambleton & Swaminathan, 1985).

The Value of Extended-Response Items

Given the nature of the current study, it is important to review how ER items potentially improve the measurement of the θ estimates for those students interested in scholarships. Of interest is previous research that indicates that ER items generally provide more information than MC items, especially for higher ability examinees (Donoghue, 1994; Wilson & Wang, 1995; Carlson, 1996). Such research suggests that ER items are better able than MC items to estimate ability and differentiate among the higher ability examinees. Further, in conjunction with polytomous IRT models, these items also provide information over a broader range of the ability (θ) scale than examinations with a similar number of dichotomously-scored items (Muraki, 1997). In contrast, Yamamoto and Kulick (1992) scaled dichotomously-scored ER items onto a polytomous scale and did not find that the polytomously-scored items produced more information than the dichotomously-scored items. However, as the authors pointed out, the items they used were not intended to be polytomously scored. Thus, the increases in information associated with polytomously-scored items may only be realized if appropriate scoring criteria are used (Samejima, 1969, 1972).

Research using items designed to be scored polytomously has shown that the ER items do provide more information. Donoghue (1994) used grades 4, 8, and 12 field data from the 1992 National Assessment of Educational Progress (NAEP) Reading assessment to examine the information produced by the dichotomously-scored MC items and the polytomously scored ER items. Separately estimating the MC items, using the three-parameter model, and the ER items, using the GPCM, Donoghue (1994) found that the ER items produced more information than the MC items. Further, the information function for the ER items peaked higher on the θ scale than the information function for the MC items.

Carlson (1996) also used the NAEP Reading data, along with NAEP data from the World Geography and United States History examinations. Unlike Donoghue, he combined both the MC and ER items to simultaneously estimate the item parameters using the three-parameter model for the MC items and the GPCM for the ER items. While he did find that the information function for the ER items for Reading did peak at a higher θ level than the MC items, such differences were not found in the other examinations. However, in the World Geography and United States History examinations, it appeared that the MC items were relatively difficult. Although Carlson did not address the issue of the information provided by the two different formats, the information functions he produced indicated that the ER items did provide more information than the MC items.

Lastly, Wilson and Wang (1995) reached conclusions similar to that of Donoghue based on their analysis of the California Learning Assessment System (CLAS) grade 4 Mathematics examination. The methods used by Wilson and Wang differed from the previous approaches in that along with simultaneous estimation of the MC and ER items using a one-parameter random coefficients multinominal logit model (Adams & Wilson, 1992), they used a "projection like approach" in which one format was used as collateral information to improve the estimates of the other format. Although limited by the use of the one-parameter model, they concluded that regardless of the approach used, the ER items provided more information and this information peaked higher on the θ scale than the MC items.

The Combined Use of Multiple-Choice (MC) and Extended-Response (ER) Items

Theoretical and technical issues arise when examinations contain both MC and ER items. From a theoretical perspective, it would seem difficult to support the simultaneous estimation of both the MC and ER items especially since the justification for the use of different types of ER items in large scale testing has largely been based on the belief that these items are measuring different traits than the MC items (e.g., Wiggins, 1993; Stiggins, 1997). Luecht (1994) has argued that combining MC and ER items to produce a single score is problematic on the basis of validity. From a psychometric perspective, these differences could have an impact on the results obtained by the methods used to determine student scores, especially when unidimensional IRT models are used (Luecht & Miller, 1992). In particular, if the two formats measure related but different traits on an examination, the assumption of unidimensionality will be violated.

Differences have been found between the two formats suggesting that the MC and ER items are measuring different traits (e.g., Thissen, et al., 1994). Bridgeman (1992) found that if the MC items from the Graduate Record Examination were transformed into dichotomously-scored ER items, the shape of some of the item response functions changed while others did not. Further, the changes were more apparent at the lower end of the θ scale. However, at the overall examination level, the results from both formats were highly correlated. In another study using the 1988 Advanced Placement (AP) Computer Science and Chemistry examinations, which contained both MC and ER items, separate factors associated with the ER items were found in addition to a larger general factor (Thissen et al., 1994). Nonetheless, since the ER factors were small as compared to the size of the general factor, the authors concluded the same trait was being measured by all of the MC and ER items. Wainer and Thissen (1993) have gone so far as to argue that even if MC and ER items measure different traits, the correlations are so high and the scoring of ER items so problematic that the use of MC items alone provides a better measure of the trait supposedly being measured by the ER items.

Recent research has compared the simultaneous and separate estimation of both the MC and ER items (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998). Based on grades 3, 5, and 8 examinations in Reading, Language, Mathematics, and Science, the authors compared the item parameters and information functions for examinations having MC and ER items combined, MC items alone, and ER items alone. In terms of model data fit, the combined MC and ER examinations did not exhibit any model data fit problems. The only difference that was found was that there was generally a slight loss of information for the ER items when they were calibrated simultaneously with the MC items rather than alone. While some of the MC-ER combinations did vary from the separate calibrations, the variations were attributed to poor reliability due to difficulty or a small number of items in the ER section. The authors concluded that the MC and ER items could and should be combined because the separate calibrations of the two formats led to scoring inconsistencies and the longer test length of the examinations having both MC and ER items would naturally increase measurement precision.

The seriousness of the unidimensionality issue is dependent on the difference in the traits being measured and the number of items measuring each trait. As the traits become more correlated, they also become more representative of the same trait (e.g., Ackerman, 1989, De Ayala, 1994, 1995). Further, if there is only a single item measuring a different but correlated trait, its inclusion with the other items will have no effect on the overall estimation process since the IRT process will weight the item accordingly (Thissen at al., 1994). It would not be beneficial to separately estimate the item parameters or determine θ estimates for the trait being measured by a single or even a small group of items due to the high standard error that would be expected with the estimation process (Luecht, 1994).

Subtest Weighting and Auxiliary Information

Two other factors considered in the current study were the inclusion of subtest weights and auxiliary information. It is possible that these procedures will produce scholarship scores that are more closely aligned to the scores calculated using the original procedure than those calculated using the current procedure.

Subtest weighting. The MC and ER sections of the provincial examinations in British Columbia are implicitly weighted through the number of marks allocated to the MC and ER sections of the examinations (see Table 1, p. 16). The reported scores are simply a sum of the scores in each of the two sections. Consequently, since the scholarship scores are now based on the total score on each provincial examination, the weights are based more on the MC section than in the original procedure. However, because different course examinations have different total test scores as well as different marks allocated to the MC and ER sections, the relative importance of the two sections potentially varies across examinations (see Table 1).

One solution is to proportionally increase the weight of the ER section within each provincial examination such that the percentage of the total score for the MC and ER sections is equal to the percentage of the total score each section provided when the original scholarship procedure was in place. Such a solution, although previously untried, could be problematic since the ER sections of the examinations are shorter and consequently, generally less reliable than the MC section. Increasing the weight of the less reliable portion of the examination will reduce reliability (Wainer & Thissen, 1994). Another weighting option is to use a criterion score to optimally weight the two sections using regression (Lord & Novick, 1968). The difficulty with this approach is that the regression weights of each format must remain constant over time.

IRT approaches combining MC and ER actually have an optimal, albeit nonlinear, weighting system built into the estimation procedure. If MC and ER items can be estimated simultaneously within an unidimensional IRT framework, the parameters for both types of items are on the same scale and optimal score "weights" are produced that are functions of the item's relation to the construct and its reliability (Wainer & Thissen, 1993). As discussed previously, the assumption of unidimensionality must first be met to realize such advantages of the IRT models. If not, the two sections must be estimated separately and weighted using the same procedures as used for the non IRT approaches.

Auxiliary Information

The use of auxiliary information to improve item parameter and θ estimates has been suggested but not widely practiced. Mislevy (1987) first suggested that the inclusion of auxiliary examinee information could improve the estimation process when using IRT models. At the time, Mislevy used auxiliary information to define the prior examinee distribution for use with an empirical Bayesian approach. However, an alternative method is to convert the auxiliary information into an examination item and estimate it along with the other items using the more commonly used maximum marginal likelihood approaches. Since the auxiliary information is free and easy to obtain in many testing situations, it could provide additional precision at a relatively low cost. Further, Mislevy (1987) was able to show that the lower bounds to the estimation of item parameters in the presence of auxiliary information was equal to the values obtained in the absence of this information. In other words, parameter estimates obtained in the presence of auxiliary information were equal to or greater than the estimates obtained in the absence of auxiliary information. While the increased precision derived from auxiliary information was related to the correlation between the information and the examination itself, Mislevy (1987) surmised that in most educational examples, this gain would translate to between two and six additional examination items. Thus, the impact of the auxiliary information would be of more value in shorter than longer examinations.

From a policy point of view, Mislevy (1987) admitted that while the estimated scores using auxiliary information would be better from a measurement perspective, the use of such information in contest or selection examinations that compared individuals would be more problematic. Since this information would likely be somewhat different than the construct being measured by the examination, it could lead to decisions based on irrelevant information. Further, such information could possibly be manipulated in order to improve the chances of specific individuals. For example, within the context of the current study, the addition of school-based mark (SBM) could unfairly increase the scholarship scores given to those students who receive higher school-based marks that are based on other factors, such as neatness, homework, or attendance, besides achievement. Nonetheless, until now, the value and effect of auxiliary information used alongside the estimation of achievement on examinations has not been carefully explored.

CHAPTER 3

METHODS

The Data Set

Given that the scholarship component was dropped during the 1996/97 school year, data for the study were obtained for the January and June examinations from the two previous years, 1994/95 and 1995/96. The sample of student responses on these examinations was large enough to examine the implications of the change in the scholarship procedure due to the deletion of the scholarship examinations. Further, by considering two years, it was possible to replicate the analyses and determine, at least partially, the stability of the findings.

The examinations analysed in the present study are listed in Table 2 together with the number of students who wrote each provincial (Prov.) and scholarship (Schol.) examination in either January or June of 1994/95 and 1995/96. Note that in the case of Geology, the provincial examination is only written in June.

Table 2

Number of Students Writing Provincial and Scholarship Examinations

	January 1995		June 1995		January 1996		June 1996	
	Prov.	Schol.	Prov.	Schol.	Prov.	Schol.	Prov.	Schol.
Biology 12	3575	1587	8583	3227	3967	1868	9114	3604
Chemistry 12	2895	1534	7908	3913	3260	1832	8003	4170
Geography 12	2496	854	6839	1717	2962	1052	6553	1903
Geology 12	--	--	1318	306	--	--	1361	308
Math 12	5342	2259	12449	5286	6147	2766	12376	5252
Physics 12	1399	772	5116	2600	1572	946	5401	2860

The Data Management Division of the British Columbia provincial government supplied the student level data necessary for the analysis. Before obtaining the data, it was necessary to obtain permission from the British Columbia Ministry of Education. Permission was obtained by completing a Research Agreement for the Disclosure of Personal Information for a Research Purpose under section 35 of the Freedom of Information and Protection of Privacy Act. This signed agreement was forwarded and subsequently approved by the Assistant Director, Information and Privacy, Finances and Administrative Services Branch of the Ministry of Education in Victoria, British Columbia. As part of the agreement, student identity was kept confidential and no student names were provided. Student identification numbers linked all student files. Computer files were kept on a single computer and password protected within a locked room thus preventing access by unauthorized personnel. Similarly, all paper files containing student identification codes were kept in a locked cabinet and shredded when no longer required. At the completion of the study, all data containing student identification codes were removed from the computer and kept on a single data CD kept by the author. After one year, this CD was destroyed. For each examination, the data included, student identification codes, gender, citizenship status, the student response vectors for the MC and ER items on the provincial examinations, provincial examination scores, school-based marks (SBM), and the original scholarship scores obtained by the students who wrote one or more scholarship examinations.

The data were analysed in seven stages. Stage 1, Data Integrity, was used to replicate the original provincial and scholarship examination results. During the second stage, Scholarship Score Calculation using the Current Procedure, scholarship scores were calculated using the current procedure. As indicated earlier, the current procedure only uses the provincial examination to determine scholarship scores. The purpose of stage 3, Examination of the Accuracy of the Current Procedure, was to compare the results of stages one and two in order to assess the potential problems of the current procedure and provide justification for the consideration of alternative procedures. The first alternative procedure considered was to replace the classical test score procedure with the generalized partial credit model (GPCM). During the next two stages, the fit of the GPCM with the provincial examinations was analysed followed by the calculation of scholarship scores using this procedure and the subsequent comparisons to the original and current procedures. During Stage 6, Scholarship Score Determination using Subtest Weights, differential weighting of the MC and the ER sections of the provincial examinations combined with the total test score model and the GPCM were implemented to calculate scholarship scores. Finally during the seventh stage, SBM was introduced as another examination item and used along with either the total test score model or the GPCM to calculate the scholarship scores.

Stage 1: Data Integrity

This stage of the analysis was completed to ensure data integrity. Using the published provincial examination keys, the student responses on the provincial examinations were rescored using S.P.S.S. 9.0 (1999) and checked against the provincial records to ensure that the individual provincial examination results were replicated. Using the procedures used by the British Columbia Ministry of Education in 1994/95 and 1995/96, the raw scores from both the provincial and the scholarship examinations for each subject and administration were summed and then rounded to the nearest whole number. The raw scholarship scores were ranked, creating a normally distribution, using the RANKIT procedure in S.P.S.S. 9.0 (see also Chambers, Cleveland, Kleiner, & Tukey, 1983). The distribution of the "ranked" scores was then standardized and transformed onto a score metric having a mean of 500 and a standard deviation of 100. The scholarship score (S_{ij}) for examinee *i* on examination *j*, is determined by:

$$S_{ij} = \left(\frac{X_{ij} - \overline{X}_{.j}}{\sigma_{.j}} \right) 100 + 500, \qquad (7)$$

where X_{ij} is the scholarship score rank for student *i* on examination *j*,

$\overline{X}_{.j}$ is the mean scholarship rank on examination *j* for the subset of students that are eligible for scholarship scores, and

$\sigma_{.j}$ is the standard deviation of the ranks on examination *j*.

Scores below 200 were set to 200 and scores above 800 were set to 800. In examinations in which the largest scholarship examination scholarship score was less than 800, the Kozlow formula was used to adjust the scores above 675 such that the highest score would be 800 (see equation 1, p. 22). All scores were rounded to the nearest whole number. The resulting scholarship scores were then compared to the scholarship scores reported by the Ministry of Education. Total scholarship scores were calculated for those students who had a minimum of three scholarship scores of 475 or more using the original procedure and those students having a total score of at least 1700 were considered scholarship recipients based on the original procedure.

Stage 2: Scholarship Score Calculation Using the Current Procedure

During this stage of the analysis, student scholarship scores were calculated using the current procedure. First, for each examination, students having a reported provincial examination score below 70% were removed from the file. In order to determine the scholarship scores for the remaining students, their reported provincial examination scores were first ranked using the RANKIT procedure in S.P.S.S. 9.0. These ranked scores were then transformed using equation 7 such that the examination scholarship mean score was 500 and the standard deviation was 100. Scores having a transformed score above 800 were adjusted to 800. If the maximum scholarship score in an examination was less than 800, the Kozlow corollary formula was used to make further adjustments to the scores (see p. 22). Again, the scholarship scores were rounded to the nearest whole number.

Finally, based on the examinations included in the current study, the results for those students who had a minimum of three scholarship scores from either the original or current procedures were kept for further analysis. As with the original procedure, the total scholarship score was calculated by summing the three highest scholarship scores for those students having three scholarship scores of 475 or more based on the current procedure. A student with a scholarship score of 1700 or more was considered to have received a scholarship while a student with a scholarship score below 1700 was considered not to have received a scholarship.

Stage 3: Examination of the Accuracy of the Current Procedure

In order to examine the extent of the changes in the scholarship scores and decisions that occurred due to the change in procedure, the results obtained using the original (stage 1) and current (stage 2) procedures were compared in terms of examination scholarship scores, total scholarship scores, and scholarship decisions. For each course examination, the comparability of the current procedure to the original procedure was examined using three different methods. First, the correlation between the scores calculated using the original and the current procedures was determined. Higher correlations indicated that the rankings of the students did not vary with the procedure used to determine scholarship scores. The correlations were expected to be high because the provincial examination scores were an integral component in both the original and current procedures. However, correlations represent a degree in similarity in ranking rather than agreement between actual scores. Therefore, the root mean square error (RMSE) was also used to examine the differences between the procedures. RMSE is a statistic used to determine the degree of agreement between two sets of scores, with lower RMSE values indicating agreement between the two sets of scores. The RMSE_j value is defined by:

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - y_{ij})^2}{n}}, \tag{8}$$

where i is a student with a scholarship score from the two procedures being compared,

n is the number of students who have scholarship scores,

x_{ij} is the scholarship score for student i on examination j using the current procedure, and

y_{ij} is the scholarship score for student i on examination j using the original procedure.

Together, the two techniques, correlations and RMSE, were used to examine if certain examination scholarship results were more affected by the change in policy.

For each examination, classification consistency was measured using a two-by-three confusion matrix as shown on Table 3. Since the writing of a scholarship examination was voluntary, only those students who attempted the scholarship examination were included in each confusion matrix. Examination scholarship scores from the original procedure were classified under two different categories, "Below 475" and "475 or higher." Three categories were required for the current procedure. A classification of "Did not qualify" indicated that a student's provincial examination score was less than 70%. A classification of "Below 475" indicated that a student's provincial examination score was at least 70% but the corresponding examination scholarship score was under 475. Finally, a classification of "475 or Higher" indicated a student's provincial examination score was at least 70% and the corresponding examination scholarship score was 475 or more.

Table 3

Sample Confusion Matrix for Examination Scholarship Scores Classification Consistency

		Scholarship Classifications Using Current Procedure			
		Did Not qualify	Below 475	475 or higher	Row totals
Original Scholarship Classifications	Below 475	Cell 1	Cell 2	Cell 3	
	475 or higher	Cell 4	Cell 5	Cell 6	
	Column totals				

In this matrix, correct classifications were for those students who were represented in cells 1, 2, or 6. Incorrect classifications were in cells 3, 4, and 5. Cell 3 signified a false positive classification in which a student's scholarship score qualified under the current procedure but failed to qualify using the original procedure. Cells 4 and 5 signified false negative classifications in which a student's score was at least 475 using the original procedure, but would either not receive a scholarship score or the score would be under 475 using the current procedure.

In order to examine how the changes in procedure influenced the decisions regarding the awarding of scholarships based on the total scholarship scores, a three by three confusion matrix, shown in Table 4, was used to compare the differences between scholarships decisions made using the original (in rows) and current (in columns) procedures. In this matrix, a classification of "Did not qualify" signified that the student had an examination scholarship score from at least three of the courses but one or more of these scores was below 475, eliminating the student from consideration for a scholarship. A classification of "Not awarded scholarship" signified that the student had three examination scholarship scores of at least 475 but the total of these three scores was below the 1700 points required to receive a scholarship. A classification of "Awarded scholarship" signified the student had three examination scholarship scores of at least 475 and would receive a scholarship since the total scholarship score was at least 1700 points.

Table 4

Sample Confusion Matrix For Total Scholarship Score Decision Consistency

		Scholarship Decisions Using Current Procedure			
		Did not qualify	Not awarded scholarship	Awarded scholarship	Row totals
Original Scholarship Decisions	Did not qualify	Cell 1	Cell 2	Cell 3	
	Not awarded scholarship	Cell 4	Cell 5	Cell 6	
	Awarded scholarship	Cell 7	Cell 8	Cell 9	
	Column totals				

Students who did not meet the minimum requirement of three examination scholarship scores of at least 475 using both the original and current procedures were represented in cell 1. Students who did not meet the 475 minimum examination score in three courses using the original procedure but would have met the 475 minimum score in three examinations totaling less than 1700 using the current procedure were represented in cell 2. Cell 3 represented those students who did not meet the three subject minimum score requirement in the original procedure but would have received a scholarship using the current procedure. Cells 4 and 7 represented those students having the opposite decision profiles from cells 2 and 3, respectively. Students who met the minimum requirement of three scholarship scores of 475 totaling less than 1700 both in the original and the current procedures were represented in cell 5. Students who met the minimum requirement of three scholarship scores of 475 in both procedures but only met the minimum total scholarship score of 1700 using the current procedure were represented in cell 6. In contrast, those students who originally received a scholarship but had three scholarship scores above 475 summing to less than 1700 using the current procedure were represented in cell 8. Finally, students who would receive a scholarship using either the original or the current procedures were represented in cell 9. The values in cells 1, 2, 4, 5, and 9 represented decisions in which the original and current procedures agreed, whereas the values in cells 3, 6, 7, and 8 represented decisions in which the two procedures disagreed. Cells 3 and 6 represented false positive decisions since students would have received a scholarship using the current procedure although they originally did not receive a scholarship. In contrast, cells 7 and 8 represented false negative decisions since students would be denied a scholarship if the current procedure was used although they actually did receive a scholarship using the original procedure.

Confusion matrices were also produced for the overall scholarship decisions for subsets of students who had a specific examination or combination of examinations as part of their total scholarship score. This was completed to determine if differences between the two procedures could be attributed to individual examinations. Further, a scholarship decision confusion matrix was produced for both males and females in order to determine if the current procedure differentially impacted males and females.

Stage 4: Scholarship Score Calculation Using the GPCM

To the extent that differences in the scholarship scores and decisions exist between the original and current procedures, alternative analytical methods that may alleviate these differences must be examined. The purposes of the analyses performed in stage 4 were to first determine if the use of the GPCM was warranted with the provincial examination data and, if so, to estimate scholarship scores using this model.

Underlying Assumptions of IRT

In order to justify the use of the GPCM, it is necessary to show that the items within an examination are suitable for analysis using the model. As summarized by Hambleton and Swaminathan (1985), the advantages of IRT can be realized only if there is a good fit between the examination data and the item response model being used. An examination of the model assumptions, in particular unidimensionality, local item independence, nonspeededness, and lack of guessing, was used to provide evidence of the fit of the GPCM model with the examination data.

Unidimensionality. An assumption of the GPCM is that the set of items within an examination provides a measure of a single (unidimensional) underlying trait. For the justification of the use of the GPCM, the assumption of essential unidimensionality is met if a single dominant component is found. For each provincial examination, unidimensionality was first examined using factor analysis. Both the MC and ER items were analysed in order to determine if they constituted a unidimensional structure. An examination was considered unidimensional if 1) the ratio of the first to second eigenvalues was considerably larger than one (in the range of five times) and the ratios of other pairs of eigenvalues were close to one and 2) the proportion of variance accounted for by the first factor was at least 15%. Since Reckase (1979) recommended that the first factor account for 20% of the total variance but obtained good θ estimates if only 10% of the variance was accounted for by the first factor, 15% was chosen since it is the midpoint between the minimum and recommended values. Scree plots were used to graphically represent the dimensionality of the data (Gorsuch, 1983). Examinations in which factor analysis did not provide clear evidence of unidimensionality were further analysed by first examining the factor structure if two factors were defined to be present and second through the use of image analysis with varimax (Gorsuch, 1983). The failure to find unidimensionality in an examination was used to examine the effect of multidimensionality on the accuracy of scholarship score estimation when unidimensional IRT models were used. For each provincial examination, the process was repeated with the SBM added as an additional item.

Local item independence. A finding that an examination was unidimensional also provided the evidence of local independence. When a single dominant trait is found to be responsible for test performance, the assumption of local independence is met (Hambleton, Swaminathan, & Rogers, 1991).

Nonspeededness. In developing the provincial examinations, efforts are made to ensure the length of the examination is appropriate for the time allocated such that students have time to attempt all questions. In other words, the examinations are designed to be power tests. Since the use of IRT models is based on the use of power tests rather than speeded tests it is important to determine if speed was a factor on the examinations. Unfortunately, since the provincial examinations consist of both MC and ER items and students are not obligated to complete one section before the other, it is more difficult to determine which items are the final items completed by the students. Thus, nonspeededness was measured by examining the omission rates on the last three MC items and the last two ER items. The assumption was considered to be met if 90% of the students completed the last three MC items and the last two ER items on a particular examination (see Hambleton and Swaminathan, 1985). If fewer than 90% of the students completed these items for a given examination, the omission rate of other items was measured to determine if the omission of the last items was likely due to time constraints or difficulty. If similar omission rates were found across items regardless of location, then the omissions were considered to be due to difficulty rather than lack of time.

Lack of guessing. Minimal guessing was only examined for the dichotomously-scored MC items. If guessing is present, the fit of the two-parameter model with respect to the data is weakened. Guessing was examined by looking at the item level performance of the students in the lowest decile on the three hardest multiple-choice items as defined by those items having the lowest *p*-values. If the performance of this subset of students on these items was similar to that of other students and was also close to 0.25 then the lack of guessing could not be assumed. If guessing was found to be a likely factor, the three-parameter model was also used for the analysis of the MC items; if not, then the two-parameter model was only used.

Calculation of the Scholarship Scores

Using the provincial examination response vectors, PARSCALE 3.1 (Muraki & Bock, 1997) was used to obtain parameter and ability estimates. Item parameter estimation was completed using the Marginal Maximum Likelihood-Expected Maximization (MML-EM) procedure with the prior ability distribution assumed to be normal (Bock & Aitkin, 1981). MML-EM is a two step iterative mathematical method. A maximum of 200 cycles was specified with a stopping criterion of 0.01 based on 30 quadrature points. The MC items were analysed using either the two- or three-parameter dichotomous model, depending on fit, and the ER items were analysed using the GPCM. If the two-parameter model was specified for the MC items, both the MC and the ER items were estimated simultaneously. Due to limitations in the PARSCALE 3.1 program, if the three-parameter model was specified for the MC items, the MC and the ER items were estimated separately and analysed in stage 6. Further, if the three-parameter model was used, the initial *c*-parameter values were set to 0.10 rather than the default setting of zero as used in PARSCALE 3.1 or the commonly used setting of 0.25. This value was used to reflect that the students writing provincial examinations are typically more academically oriented than the population of students as a whole. The student θ estimates were determined using the *expected a posteriori* (EAP) estimator, a Bayes estimate that is the mean of the posterior distribution of θ given the observed response pattern (Bock &

Mislevy, 1982). The EAP estimate works for all response patterns (including perfect scores) and has a smaller average error than any other method (Bock & Mislevy, 1982).

For each examination, the initial estimation was completed using the population of students who completed the examination. After the θ estimates were determined for all of the students, those with a reported provincial examination score of less than 70% were dropped from further analysis. The θ estimates of the remaining subset of students were transformed into examination scholarship scores using the same methods as used for the previous procedures. For the sake of comparisons during the study, this procedure was called the GPCM procedure (GPCM).

Stage 5: Examination of the Accuracy of the GPCM

The results of the GPCM and the original procedures were compared using the same methods that were used to compare the original and current procedures. Additionally, the scholarship scores, classifications, and decisions obtained using the GPCM and the current procedures were compared to determine if the use of GPCM was either comparable or superior to the current procedure.

Stage 6: Scholarship Score Determination using Subtest Weighting

Given the research regarding the value of ER items and since the scholarship examinations consisted entirely of ER items, a weighting approach with differential weights on the MC and the ER sections of each examination was completed in an attempt to produce scholarship scores and decisions using only the provincial examinations that better replicated the original procedure (Donoghue, 1994; Wilson & Wang, 1995; Carlson, 1996). Further, the weighting was combined with both the classical (total test score) and IRT models. In order to complete these analyses, the MC and the ER items of each provincial examination were separated into two subtests. In the case of the total test score model, a total score for each subtest was calculated. In the case of the IRT models, the two subtests were estimated separately in PARSCALE 3.1 to produce two θ estimates. Separate analyses were also done using both the two- and three-parameter models for the MC items. Two alternative weighting procedures were examined, scholarship weighting and optimal weighting, in combination with the three different models above, for a total of six different procedures.

In order to obtain scholarship weighting for each examination, the student scores or θ estimates from the MC and ER subtests were weighted such that the contribution of each mirrored the contribution of the MC and ER sections to the examination scholarship scores when both the provincial and scholarship examinations were used to determine these scores. Since each scholarship examination only contained ER items, the contribution of the ER subtest was increased while the contribution of the MC subtest was decreased. Three scholarship weighting procedures were used, Classical, (Scholarship), GPCM, (2-parm; Scholarship), and GPCM, (3-parm; Scholarship).

In order to obtain optimal weights, stepwise regression was used to determine the weights of the MC and ER sections that best predicted the original raw scholarship scores (sum of the provincial and scholarship score) for each course. Furthermore, this regression was based only on those students that had an original examination scholarship score of at least 475. This ensured that the regression values were based only on those students for whom scholarship scores would have qualified as part of their total scholarship score. Three optimal weighting procedures were used, Classical, (Optimal), GPCM, (2-parm; Optimal), and GPCM, (3-parm; Optimal).

Once the weighted scores or θ estimates were determined, the calculation of scholarship scores used the procedures described in stages 2 and 4 above. Comparisons to the original examination and total scholarship scores were made using correlations, RMSEs, and confusion matrices. The results were also compared to those of the current and other alternative procedures to determine if the procedures using subtest weighting better replicated the original scholarship scores and decisions than the other procedures.

Stage 7: Scholarship Score Determination using Auxiliary Information

During stage 7, auxiliary information was used to create another examination item based on the each student's school-based mark (SBM). The SBM was transformed onto the six-point scale as summarized in Table 5. In the case of the total test score model, three different procedures were attempted using SBM. First, each student's SBM score was added to his/her provincial examination score as another item worth 10 percent of the total test score (Classical, MC-ER-SBM). Second, the provincial examination and the SBM were treated as two subtests and optimally weighted using the method described in stage 6 (Classical, SBM (Optimal)). Lastly, the MC, ER, and SBM were treated as three separate subtests and optimally weighted using the method described in stage 6 (Classical, MC/ER/SBM (Optimal)).

Table 5

Guidelines for School-Based Mark (SBM) Transformation

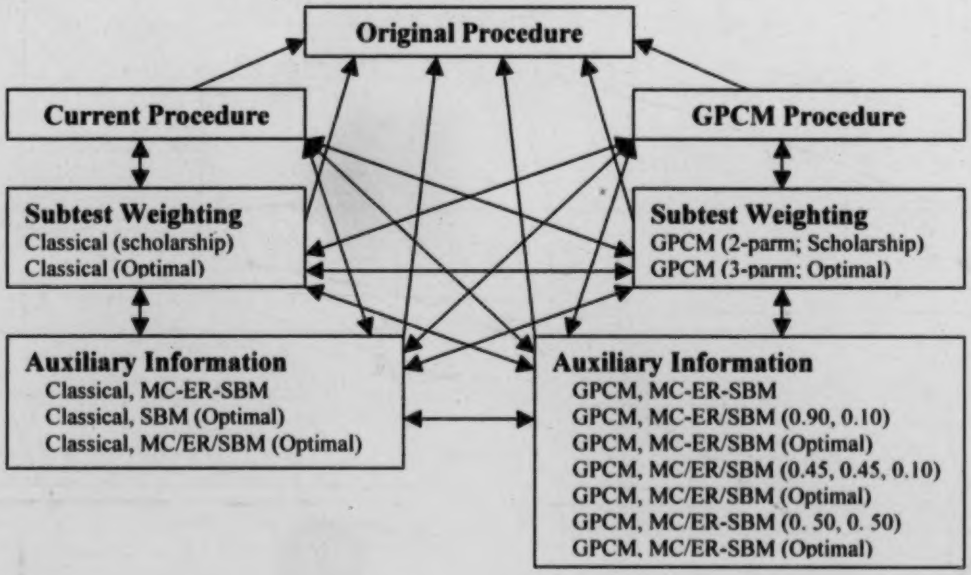
SBM (%)	SBM Polytomous Score
0 to 49	1
50 to 59	2
60 to 69	3
70 to 79	4
80 to 89	5
90 to 100	6

In the case of the GPCM, seven different procedures were attempted. First, all of the items from the provincial examination were estimated simultaneously with SBM (GPCM, MC-ER-SBM). Second, the MC and ER items were estimated simultaneously as one subtest and the SBM as a second subtest. These two θ estimates were then weighted 0.90 and 0.10, respectively (GPCM, MC-ER/SBM (0.9, 0.10)). This procedure was repeated with the exception that optimal regression was used to determine the contribution of the θ estimate from each subtest (GPCM, MC-ER/SBM (Optimal)). In the next two analyses, the MC, ER, and SBM were estimated separately as three separate subtests and then in the first instance weighted such that the contribution of the θ estimate from each subtest was 0.45, 0.45, and 0.10, respectively (GPCM, MC/ER/SBM (0.45, 0.45, 0.10)). In the second instance the contribution of each θ estimate was based on the optimal weights (GPCM, MC/ER/SBM (Optimal)). In the last pair of analyses, the MC items were considered one subtest while the ER items and SBM were combined to form a second subtest. Each subtest was estimated separately and the θ estimates combined, in the first instance such that the θ estimate from each subtest contributed 0.50 to the total θ estimate (GPCM, MC/ER-SBM (0. 50, 0. 50)). In the second instance, the contribution of each θ estimate was based on optimal weights (GPCM, MC/ER-SBM (Optimal)).

The calculation of the scholarship scores in this stage used the methods described previously in stages 2 and 4. Comparisons to the original examination and total scholarship scores and decisions used correlations, RMSEs, and confusion matrices. The results of the 10 procedures in this stage were also compared to those of the current and other alternative procedures to determine if the procedures using SBM better replicated the original scholarship scores and decisions than the other procedures.

Figure 4 illustrates the levels of the different procedures used and the comparisons that were made. All of the procedures were compared to the original procedure as shown by the single directional arrows. Further, the current and alternative procedures were also compared with each other as shown by the bi-directional arrows.

Figure 4. Flow Chart Illustrating the Procedures used and the Comparisons Made.


Original Procedure
Current Procedure
GPCM Procedure
Subtest Weighting
Classical (scholarship)
Classical (Optimal)
Subtest Weighting
GPCM (2-parm; Scholarship)
GPCM (3-parm; Optimal)
Auxiliary Information
Classical, MC-ER-SBM
Classical, SBM (Optimal)
Classical, MC/ER/SBM (Optimal)
Auxiliary Information
GPCM, MC-ER-SBM
GPCM, MC-ER/SBM (0.90, 0.10)
GPCM, MC-ER/SBM (Optimal)
GPCM, MC/ER/SBM (0.45, 0.45, 0.10)
GPCM, MC/ER/SBM (Optimal)
GPCM, MC/ER-SBM (0. 50, 0. 50)
GPCM, MC/ER-SBM (Optimal)

CHAPTER 4

ANALYSIS OF THE CURRENT PROCEDURE

The results of the analyses conducted to address the question regarding the accuracy of the current procedure are addressed in this chapter. The chapter is organized in three sections corresponding to the three stages of the analyses that focused on the current procedure. Within each section, the results for the 1994/95 school year are presented first while the results for the 1995/96 school year are presented second. Section 1, Data Integrity, contains the results that were obtained by comparing the provincial examination and scholarship scores as determined by the Ministry of Education in 1994/95 and 1995/96 with the provincial examination and scholarship scores as calculated from the raw data used for the current study. Section 2, Scholarship Score Calculation using the Current Procedure, contains the descriptive information regarding the number of students who wrote scholarship examinations or qualified using both the original and current procedures as well as the distribution of scholarship scores for these procedures. Section 3, Examination of the Accuracy of the Current Procedure, contains the results obtained in comparing the original and current scholarship scores and decisions. In examining the current procedure, the original scholarship results were considered the correct scores and decisions. Deviations from the original scholarship scores or decisions that would occur if the current procedure was used in place of the original procedure provided evidence of the extent of the error and incorrect decisions made using the current procedure.

Section 1: Data Integrity

The purpose of this stage of the analysis was to ensure the data that arrived from the British Columbia Ministry of Education was the same as that used during the 1994/95 and 1995/96 school years. For all of the courses, provincial examination scores derived from the raw data produced the same raw provincial examination scores as those reported by the Ministry of Education in those years. Similarly, after using the methods described in Chapter 3 (see p. 52) to calculate and adjust the scholarship scores, the examination scholarship scores calculated using the original procedure were the same as those reported by the Ministry of Education in 1994/95 and 1995/96 within one score point. This difference of one scholarship point occurred rarely and likely was due to a difference in the rounding procedure used. It was therefore concluded that the data for the two years analysed in the present study were the same as those used by the Ministry of Education.

Section 2: Scholarship Score Calculation Using the Current Procedure

Using the current procedure, students with provincial examination scores of 70% or more receive a scholarship score. For each examination analysed, the scholarship score transformation and correction methods described in Chapter 3 (see p. 53) were used to produce examination scholarship scores for these students. The descriptive information for the examination scholarship scores derived from both the original and the current procedures are presented in Table 6. Table 6 is separated into four panels. The two left panels contain the means and standard deviations for the examination scholarship scores derived first from the original procedure and then from the current procedure for the 1994/95 school year. The two right panels contain the same information for 1995/96.

Table 6

Descriptive Statistics for Scholarship Scores using the Original and Current Procedures

	1994/1995				1995/1996			
	Original		Current		Original		Current	
Subject	M	SD	M	SD	M	SD	M	SD
Biology (January)	499.62	98.81	500.40	98.30	499.55	98.78	500.52	97.87
Biology (June)	499.69	98.70	500.78	97.33	499.71	98.95	500.63	97.50
Chemistry (January)	499.85	99.14	500.45	99.01	499.79	99.55	502.08	103.04
Chemistry (June)	499.63	98.74	500.53	98.57	499.61	98.69	500.77	100.14
Geography (January)	499.80	99.07	500.81	97.48	499.70	98.99	500.83	97.38
Geography (June)	499.92	99.40	501.73	95.21	499.52	99.16	501.31	96.33
Geology (June)	499.96	99.90	501.09	96.81	500.03	100.03	501.43	97.92
Mathematics (January)	499.98	98.58	501.05	99.24	500.00	99.49	500.85	99.52
Mathematics (June)	500.23	100.48	502.06	51.36	499.75	99.29	501.08	100.51
Physics (January)	499.72	99.18	500.72	99.73	499.77	99.03	500.13	98.83
Physics (June)	500.14	100.23	501.06	101.20	499.63	98.85	500.50	98.83

Note: Sample sizes for each examination are provided on Table 7, p. 74.
These descriptives are based on the final adjusted scholarship scores (based on the Kozlow Correction formula, p. 22) for each examination.

The means for the scholarship scores using the current procedure were slightly higher, to a maximum of 2.53 (January 1996, Chemistry), than the means for the original procedure while the standard deviations for the current procedure were slightly less, to a maximum of 4.19 (June 1995 Geography). These differences were likely due to the decreased raw score range using the current procedure. This would decrease the score variation and subsequently increase the use of the Kozlow formula (see p. 22). This adjustment of the highest scores would increase the means but have less effect on the standard deviations.

When the original procedure was in place, students expressed their desire to try to obtain a scholarship by writing the optional scholarship examinations. With the current procedure, all students who have a provincial examination score of at least 70% are automatically given a corresponding scholarship score. Thus, there are some differences not only in the number but also the population of students receiving scholarship scores between the two procedures. In order to compare the original and current procedures only the subset of students that met the scholarship requirements under both procedures could be meaningfully compared. The number of students who had scholarship scores under the original and current procedures as well as the sample of students who qualified using both procedures is illustrated in the Venn diagram in Figure 5. In this figure, the numbers not in parenthesis are for 1994/95 while the numbers enclosed in parenthesis are for 1995/96. As illustrated, there were 12,121 students who wrote at least one scholarship examination in the analysed courses during the 1994/95 year. Of this number, 3,525 students wrote at least three scholarship examinations in these courses. Using the current procedure, 14,133 students received a provincial examination score of at least 70% in at least one of the courses considered. Of this number, 3,319 students met the 70% criteria in at least three of these courses. The overlap between the students who wrote at least three scholarship examinations and achieved at least 70% on three or more provincial examinations was 2,524.

For 1995/96, 13,422 students wrote at least one of the scholarship examinations in the analysed courses and 3,874 students wrote at least three scholarship examinations. Using the current procedure, 15,788 students achieved a provincial examination score of at least 70%. Of this number, 3,633 students met the 70% criteria in at least three of the courses. Together, 2,769 students wrote at least three scholarship examinations and achieved a minimum of 70% on three or more provincial examinations.

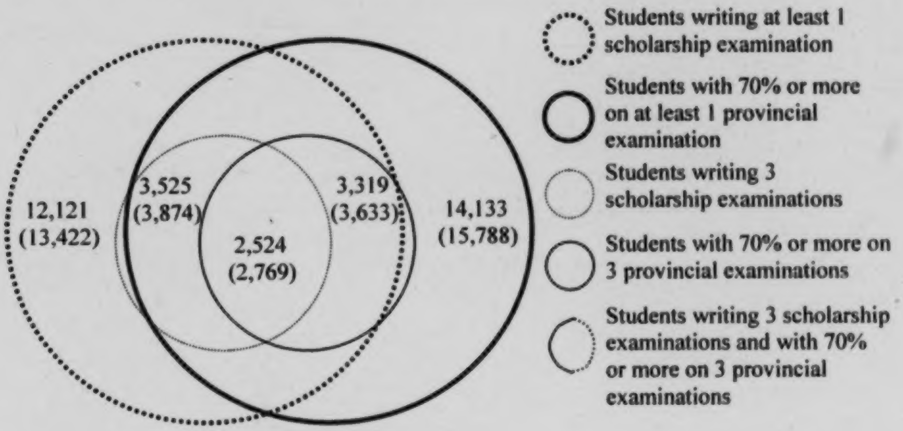

12,121
(13,422)
3,525
(3,874)
2,524
(2,769)
3,319
(3,633)
14,133
(15,788)
Students writing at least 1 scholarship examination
Students with 70% or more on at least 1 provincial examination
Students writing 3 scholarship examinations
Students with 70% or more on 3 provincial examinations
Students writing 3 scholarship examinations and with 70% or more on 3 provincial examinations

Figure 5. Students with Scholarship Scores from either the Original or the Current Procedures

Section 3: Examination of the Accuracy of the Current Procedure

The comparability of the original and the current procedures was examined both at the examination and the total scholarship score levels through the similarity in the identification of the students receiving scholarship scores, correlations, RMSE values, and decision consistencies. Confusion matrices were used to examine decision consistency at the examination and total scholarship score levels. Further, total scholarship score confusion matrices were produced for subsets of students having specific examinations or examination combinations as part of their total score as well as for gender (see Chapter 3, p. 59).

Individual Examination Scholarship Scores

The consistency and agreement between scholarship scores as calculated by the original and current procedures was examined to determine if differences in the scholarship scores could be attributed to differences between the two procedures at the individual examination level. These comparisons are summarized in Table 7. Like Table 6, this table is separated into two panels. The left panel includes those students who received a scholarship score for each course examination during the 1994/95 school year and the right panel for 1995/96. First, the two procedures were compared in terms of the students who qualified using each procedure. For example, as shown in the second column, 1,587 students wrote the Biology scholarship examination in January 1995. Had the current procedure been employed, 1,576 (column 3) students who wrote this Biology provincial examination would have received a scholarship score. The number of students who would have received a corresponding examination scholarship score for the January 1995 sitting of the Biology examination using both of the procedures was 1,155 (column 4). Column 5 gives the correlation between the original and the current procedure for this latter sample. For the January 1995 Biology examination, this correlation was 0.91. Similarly, as shown on column 6, the corresponding root mean square error (RMSE) between the two procedures was 48.26 or 8.0% of the scholarship score range of 600 (200-800).

Table 7

Comparison of Scholarship Results Between the Original and Current Procedures

	1994/1995					1995/1996				
Subject	Original	Current	Overlap	r	RMSE	Original	Current	Overlap	r	RMSE
Biology (January)	1587	1576	1155	0.91	48.26	1867	2005	1488	0.92	40.80
Biology (June)	3227	3751	2533	0.91	42.60	3604	4310	2927	0.92	38.68
Chemistry (January)	1534	1315	1045	0.91	58.42	1833	1961	1481	0.90	47.50
Chemistry (June)	3913	4149	3109	0.92	44.01	4170	4386	3309	0.91	45.67
Geography (January)	854	970	544	0.74	71.69	1052	1042	660	0.71	74.03
Geography (June)	1717	2386	1196	0.76	64.40	1904	2763	1384	0.77	62.94
Geology (June)	306	422	219	0.81	57.04	308	344	197	0.85	64.40
Mathematics (January)	2259	2305	1652	0.90	51.43	2765	2553	1934	0.87	55.33
Mathematics (June)	5283	5322	3893	0.90	51.36	5252	5423	3892	0.88	52.67
Physics (January)	772	660	533	0.88	64.34	945	852	705	0.88	55.84
Physics (June)	2601	2899	2118	0.88	48.40	2860	2873	2201	0.89	47.49

Examination of the results in Table 7 reveals there were differences between the examination scholarship scores calculated using the two procedures. First, the number of students who would receive scholarship scores using the two procedures varied somewhat. With six exceptions, the number of students who wrote the scholarship examination was less than the number of students who would have received a scholarship score using the current procedure. The six exceptions were all examinations administered in January. It may be that some students, aware that they did poorly on the January scholarship examinations they wrote, may have decided not to write any scholarship examinations in June. Potentially, this could imply the pool of students who wrote scholarship examinations in June was different than the pool of students who wrote in January. Second, the students receiving scholarship scores within each procedure differed. For example, for the January 1995 Biology examination, 853 students (1587 – 1155 + 1576 – 1155) would qualify under only one of the two procedures (row 1, Table 7). Again, part of this difference could be attributed to the optional nature associated with the original procedure in which students elected to write the scholarship examination. Nevertheless, the discrepancy does signify a potential problem in the equality of the two procedures.

As with the number of students obtaining scholarship scores under either procedure, the correlations between the two procedures provide some evidence regarding their comparability. With the exception of Geography and the 1995 Geology examinations, the correlations between the scores yielded by the two procedures indicated that the results were stable across the two years (between 0.85 to 0.92). While these correlations are moderately high, the values indicate that the scholarship examinations did have an effect on the ranking of students with respect to their examination scholarship scores. The correlations in Geography, and to a lesser extent Geology, were lower, suggesting that the removal of these scholarship examinations resulted in greater changes in the ranking of the students in these two courses. Again, these differences suggest a potential problem with the current procedure.

As with the student numbers and the correlations, the average differences between the scholarship scores as determined using the original and current procedures signify potential problems. Given that the range of scholarship scores is 600 points, the RMSE value of 48.26 for the January 1995 Biology examination represents an average score error of 8.0% of the range. Across the four examination periods, the RMSE values varied from 38.68 (6.4%) to 74.03 (12.3%) with the average being 53.97 (9.0%). However, the RMSEs were lower for the June examinations than for the January examinations for every examination except the 1995 Mathematics examination. In the most extreme example, the RMSE between the January 1995 Physics examination was 15.94 points (2.7%) higher than the RMSE for the June 1995 Physics examination. Such differences may be due to differences in the sample of students who wrote the examinations at the two different times as mentioned above or to errors associated with the smaller number of students who wrote in January. With respect to the smaller sample size, a correlation of –0.60 (p<.01) was found between sample size and RMSE, suggesting that there was a negative relationship between the number of students writing an examination and RMSE.

Turning to the issue of classification consistency, two-by-three confusion matrices were used to compare the classifications between the original and current procedures. The confusion matrix for the June 95 Biology 12 examination is presented in Table 8 as an example. The three cells in bold represent wrong classification decisions. The 3 (0.1%) students who did not qualify and the 222 (6.9%) who obtained scholarship scores below 475 using the current procedure were considered false negative classifications because these students had examination scholarship scores of 475 or higher using the original procedure. In contrast, the 63 (2.0%) students represented in cell 3 were considered false positive classifications because these students had scores less than 475 using the original procedure but would have received scholarship scores of 475 or higher using the current procedure. Combined, the total classification error rate for the June 1995 Biology 12 examination was 8.9%.

Table 8

Confusion Matrix for the Biology 12 June 1995 Examination

		Scholarship Classification Using Current Procedure			Row totals
		Did Not Qualify	Below 475	475 or higher	
Original Scholarship Classification	Below 475	691 (21.4)	489 (15.2)	**63 (2.0)**	1243 (38.5)
	475 or higher	**3 (0.1)**	**222 (6.9)**	1759 (54.5)	1984 (61.5)
	Column totals	694 (21.5)	711 (22.0)	1822 (56.5)	3227

A summary of the numbers and percentages of false positive and false negative classifications and the total percentage error that would have occurred if the current procedure was used instead of the original procedure are presented in Table 9. Like Table 7, Table 9 is separated into two panels. The results for 1994/95 are reported in the left panel and the results for 1995/96 are reported in the right panel. For each examination, the numbers are reported along with the corresponding percentages in brackets.

Table 9

Scholarship Examination Classification Errors using the Current Procedure

Subject	1994/1995 Number of Students	False Negative	False Positive	Error Rate (%)	1995/1996 Number of Students	False Negative	False Positive	Error Rate (%)
Biology (January)	1587	167 (10.5)	8 (0.5)	11.0	1867	132 (7.1)	26 (1.4)	8.5
Biology (June)	3227	225 (7.0)	63 (2.0)	8.9	3604	204 (5.7)	87 (2.4)	8.1
Chemistry (January)	1534	239 (15.6)	3 (0.2)	15.8	1833	130 (7.1)	43 (2.3)	9.4
Chemistry (June)	3913	301 (7.7)	64 (1.6)	9.3	4170	345 (8.3)	83 (2.0)	10.3
Geography (January)	854	161 (18.9)	18 (2.1)	20.6	1052	204 (19.4)	35 (3.3)	22.7
Geography (June)	1717	258 (15.0)	59 (3.4)	18.5	1904	207 (10.9)	70 (3.7)	14.5
Geology (June)	306	30 (9.8)	11 (3.6)	13.4	308	60 (19.5)	4 (1.3)	20.8
Mathematics (January)	2259	244 (10.8)	15 (0.7)	11.5	2765	383 (13.9)	39 (1.4)	15.3
Mathematics (June)	5283	471 (8.9)	48 (0.9)	9.8	5252	542 (10.3)	83 (1.6)	11.9
Physics (January)	772	127 (16.5)	8 (1.0)	17.5	945	121 (12.8)	8 (0.8)	13.7
Physics (June)	2601	178 (6.8)	92 (3.5)	10.4	2860	332 (11.6)	62 (2.2)	13.8
Average	1850	185 (6.5)	30 (3.4)	11.3	2043	205 (7.0)	42 (3.2)	11.5

Note: Differences in the overall error rate are due to rounding
Numbers in brackets are the corresponding percentages

For the 1994/95 examinations, the total classification error rates varied from a low of 8.9% for the June 1995 Biology examination to a high of 20.6% for the January 1995 Geography examination and the mean error rate was 11.3%. For each examination, the total error rate was higher for the January examinations than for the corresponding June examinations. Further, the majority of the errors were false negative classifications with a greater proportion of false negative classifications occurring in January. The mean ratio of false negative to false positive errors was 18.0 to 1 for the January session as compared to 4.3 to 1 for the June session. This suggests the current procedure generally underestimated the number of students who should receive the minimum scholarship score in each course examination and that this problem was more pronounced in the January examinations.

A similar error profile was found for the 1995/96 examinations. The error rates varied from a low of 8.1% for the June 1996 Biology examination to a high of 22.7% for the January 1995 Geography examination with the average error rate being 11.5%. With the exception of Chemistry and Physics, the total error rate was higher during the January sitting of the examinations than during the June sitting. The majority of the errors were false negative classifications and a larger proportion of false negative classifications occurred in the January examinations. However, the ratios were more similar than the previous year. The mean false negative to false positive ratio was 6.4 to 1 for the January examinations and 4.3 to 1 for the June examinations. Nevertheless, the 1995/96 results provide further support that the current procedure generally underestimates the number of students who should receive the minimum scholarship score in each course examination and that a larger proportion of these errors occur in January.

Based on the classification consistency results for the examination scholarship scores, it appears the differences between the January and June examinations may be due to more than sample size. For example, the June 1995 Geology examination had a false negative/false positive ratio that was similar to the other June examinations although the sample was smaller than any of the January examinations. Similarly, the proportion of false negative to false positive errors on the two January Mathematics examinations was similar to the other January examinations even though the samples were similar in size to the June examinations in both Geography and Physics.

Total Scholarship Score Results

While an examination of the results at the individual examination level may provide some insight into possible differences between the original and the current procedures, the impact of the differences could only be determined through a comparison of the two procedures at the total scholarship score level. At this level, the analysis of each student's total scholarship score was based on the summed total of his/her three highest examination scholarship scores. Analysis at this level continued to indicate that there were differences between the two procedures. For example, based on Figure 5 and described previously, 3,525 and 3,874 students wrote at least three scholarship examinations while 3,329 and 3,633 students had 70% on at least three provincial examinations in 1994/95 and 1995/96, respectively. Thus, in the case of the total scholarship score, more students qualified using the original procedure as opposed to the current procedure. Interestingly, this was different than the results found at the individual examination level in which more students qualified using the current procedure. As with the examination level results, there was a substantial difference in the students who received a total scholarship score depending on the procedure used. Using the 1994/95 results given in Figure 3, 2,524 students qualified under either procedure, whereas a total of 1,806 students (3,329 – 2,524 + 3,525 – 2,524) qualified based on only one of the procedures. These results suggest that the shift to the current procedure has led to different students obtaining a total scholarship score.

The decision consistency at the total scholarship score level was determined using a three by three (original procedure-by-current procedure) confusion matrix (see Table 4, p. 57) for each year. The results for both 1994/95 and 1995/96 are presented in Table 10 with the first row in each cell containing the 1994/95 numbers and percentages (in brackets) and the second row containing the 1995/96 results. The numbers in bold print represent incorrect decisions. Of the 2,524 students who met the three examination scholarship score requirements for both the original and current procedures during the 1994/95 school year, 164 students (41 + 123) incorrectly would be denied a scholarship using the current procedure since they were given a scholarship using the original procedure (false negative). In contrast, 85 students (9 + 76) incorrectly would be given a scholarship using the current procedure because they did not meet the necessary requirements under the original procedure (false positive). In total, 249 students or 9.9% incorrectly would be denied or given a scholarship in 1994/1995 using the current procedure based on the set of examinations analysed. As with the individual examination results, the proportion of false negative decisions was larger than false positive decisions; however, at 1.9 to 1, the ratio was smaller than that found at the course level.

Of the 2,769 students with total scholarship scores from both procedures in 1995/96, 195 students incorrectly would be denied a scholarship (false negative) and 89 students incorrectly would be given a scholarship (false positive). In total, the error rate for the current procedure was 10.3%. Once again, at 2.2 to 1, the proportion of false negative decisions was larger than the proportion of false positive decisions.

Table 10

Confusion Matrix for the Scholarship Decisions Comparing the Original and the Current Procedures

		Scholarship Decisions Using Current Procedure			
		Did not qualify	Not awarded scholarship	Awarded scholarship	Row totals
Original Scholarship Decisions	Did not qualify	615 (24.4%)	47 (1.9%)	**9 (0.4%)**	671 (26.6%)
		651 (23.5%)	61 (2.2%)	**14 (0.5%)**	726 (26.2%)
	Not awarded scholarship	294 (11.6%)	381 (15.1%)	**76 (3.0%)**	751 (29.7%)
		401 (14.5%)	390 (14.1%)	**75 (2.7%)**	866 (31.3%)
	Awarded scholarship	**41 (1.6%)**	**123 (4.9%)**	938 (37.2%)	1102 (43.7%)
		58 (2.1%)	**137 (4.9%)**	982 (35.5%)	1177 (42.5%)
	Column totals	950 (37.6%)	551 (21.8%)	1023 (40.5%)	2524
		1110 (40.1%)	588 (21.2%)	1071 (38.7%)	2769

Note: Numbers in brackets represent percentages of the total
The top row of each cell is for 1994/95. The second row is for 1995/96
Numbers in bold represent incorrect decisions

Based on the comparisons between the original and current procedures, it is evident that the shift to the current procedure has indeed caused differences not only in the total scholarship scores that students would achieve but also in the awarding of scholarships. While the differences did vary at the individual examination level, the net result was an error rate in the determination of scholarship recipients of 9.9% for the 1994/95 school year and 10.3% for the 1995/96 school year. The use of the three highest examination scholarship scores to determine scholarship recipients did appear to mediate some of the errors associated with individual examinations and did lower the ratio between the false negative and false positive decisions. Across 1994/95 and 1995/96, the false negative rates were 6.5% and 7.0%, respectively, while the corresponding false positive rates were 3.4% and 3.2%, indicating that the false negative rates were approximately two times the false positive rates. Nonetheless, an overall error rate of approximately 10% across both years signifies that one in ten students would be treated unfairly by the current procedure.

Decision Consistency in Subsets of Examinations

In an attempt to clarify the causes of the decision errors associated with the current procedure, an analysis of the classification errors for particular subsets of examinations was completed. The total scholarship score decision consistency of the current procedure was examined for subsets of examinations based on the time the examinations were written, specific course examinations, and specific sets of examinations. These results are reported in Table 11. As with Table 9, the numbers are provided with the corresponding percentages in brackets. The left panel contains the results for 1994/95 and the right panel the results for 1995/96. For the purposes of comparison, the first row contains the results for the total sample as reported in Table 10. The next three rows summarize the decision consistency associated with the time of year that the examinations were written. For example, as seen in the second row, 109 students (column 2) derived their total scholarship score from examinations written only in January. Of these students, nine (8.3%) students (column 3) would have been incorrectly denied a scholarship using the current procedure and no students (column 4) would have been incorrectly given a scholarship. Thus, the total error rate for the students writing the

Table 11

Scholarship Decision Errors using the Current Procedure for Subsets of Examinations

Subject	1994/1995 Number of Students	1994/1995 False Negative	1994/1995 False Positive	1994/1995 Error Rate (%)	1995/1996 Number of Students	1995/1996 False Negative	1995/1996 False Positive	1995/1996 Error Rate (%)
Current Procedure (Overall)	2524	164 (6.5)	85 (3.4)	9.9	2769	195 (7.0)	89 (3.2)	10.3
3 January examinations	109	9 (8.3)	0 (0.0)	8.3	119	5 (4.2)	3 (2.5)	6.7
January and June exams	1061	76 (7.2)	31 (2.9)	10.1	1345	94 (7.0)	51 (3.8)	10.8
3 June examinations	1354	79 (5.8)	55 (4.1)	9.9	1305	96 (7.4)	34 (2.6)	10.0
Biology included	1282	76 (5.9)	43 (3.4)	9.3	1569	91 (5.8)	57 (3.6)	9.4
Chemistry included	2165	141 (6.5)	67 (3.1)	9.6	2353	165 (7.0)	76 (3.2)	10.2
Geography included	432	24 (5.6)	24 (5.6)	11.1	524	31 (5.9)	23 (4.4)	10.3
Geology included	54	7 (13.0)	4 (7.4)	20.4	66	8 (12.1)	2 (3.0)	15.2
Math included	2173	139 (6.4)	70 (3.2)	9.6	2283	166 (7.3)	72 (3.2)	10.4
Physics included	1472	105 (7.1)	47 (3.2)	10.3	1517	125 (8.2)	37 (2.4)	10.7
Chem./Math/Phys. (June)	592	42 (7.1)	22 (3.7)	10.8	486	46 (9.5)	9 (1.9)	11.3
Bio./Chem./Math (June)	377	15 (4.0)	15 (4.0)	8.0	387	18 (4.7)	13 (3.4)	8.0
Chem./Phys.(Jun) /Math(Jan)	146	8 (5.5)	4 (2.7)	8.2	159	15 (9.4)	7 (4.4)	13.8
Bio./Chem.(Jun) /Math(Jan)	84	4 (4.8)	1 (1.2)	6.0	101	11 (10.9)	5 (5.0)	15.8

Note: Differences in the overall error rate are due to rounding
Numbers in brackets are the corresponding percentages

subset of examinations in January 1994/95 was 8.3% (column 5). The third row contains those students who derived their total scholarship scores from both January and June examinations while the fourth row includes those students who derived their total scholarship scores only from the June examinations.

The results indicate that despite the larger error rates associated with the individual January examinations, the identification of scholarship winners using the current procedure in place of the original procedure was largely unaffected by the session the students completed their examinations. The lowest total error rate (8.3% for 1994/95 and 6.7% for 1995/96) occurred for those students who derived their total scholarship scores solely from the January examinations. While the sample size was relatively small and subject to variation, this relatively lower error rate may also be attributable to differences in the unique population of students who would be completing three provincial examinations in January. It is possible these students were those very strong academic students who were completing their grade 12 provincial examinations in January in order to prepare for and complete Advanced Placement examinations or the International Baccalaureate programs the following Spring. Unfortunately, it is not possible to examine this hypothesis more closely at this time so it remains speculation. Unlike the three January examination subset, no unique findings could be found for the other time combinations suggesting that these combinations had little differential effect on the total error rate.

The next six rows in Table 11 provide the results for students' total scholarship scores who had one of their three examination scholarship scores from each of the specific courses considered in the present study. With the exception of combinations having Geology, the individual examinations did not appear to have a differential effect on the overall error rate since the total error rate was near 10% for each. Students who had a total scholarship score that included an examination scholarship score from Geology were more likely misclassified than those who had a total scholarship score that did not include a Geology score. However, the number of students having Geology as one of their examination scholarship scores was small, thereby minimizing the effect of this higher error rate. While the total error rate for total scholarship scores containing the other examinations was similar, the distribution of the errors differed. For example, the ratio between false negative and false positive decisions was much smaller when Geography was included as one of the examinations, being 1.0 in 1994/95 and 1.3 in 1995/96. In contrast, this ratio was larger when Physics was included as one of the examinations (2.3 and 3.4). Physics is considered one of the most demanding academic courses and it attracts the most academic students while Geography is considered less demanding, thus attracting less academic students. Similar but less noticeable trends were found for the other subjects with the ratio for total scholarship scores including Biology being smaller and those including Chemistry or Mathematics being larger. If these findings are true, then there may be a slight bias with the current procedure against more academic students.

The final four rows of Table 11 include a summary of the decision consistencies for students who had total scholarship scores derived from a specific combination of three provincial examinations. The first combination, the June sessions of Chemistry, Mathematics, and Physics, was investigated because it had the largest number of students and the lowest RMSE values (see Table 7, p. 74). The second combination, the June sessions of Biology, Chemistry, and Mathematics, was of interest because it contained the three examinations with the lowest proportion of classification errors (see Table 9, p. 78). Both of these combinations had relatively consistent error rates across the two years and the Biology, Chemistry, and Mathematics combination also had a lower total error rate (8.0% for both years) than the overall error rate (see row 1). The false negative to false positive ratio was also smaller than average for this combination. The Chemistry, Mathematics, and Physics combination was slightly above the average in terms of error rate and for the 1995/96 year, the false negative to false positive ratio was quite high (5.1 to 1). Other combinations were examined but as illustrated by the final two rows on the table, the smaller sample sizes seemed to have affected the consistency of the results. Given that other combinations had even smaller samples, they were not analysed.

Decision Consistency for Males and Females

The scholarship decision consistency for males or females was examined to determine if errors in the current procedure could be attributed to gender. Based on the examinations included in the study, fewer females than males had three scholarship scores from both procedures, 1,088 versus 1,436 in 1994/95 and 1,226 versus 1,543 in 1995/96. While generally fewer females had the minimum of three scholarship scores necessary to receive a total scholarship score, the proportion of decision errors for females was lower than that for males, 7.8% versus 11.4% in 1994/1995 and 9.1% versus 11.2% in 1995/1996. However, the ratio of false negative to false positive decisions was virtually the same for both genders being approximately two to one. This is surprising since males have been reported to have greater performance on multiple-choice (MC) items and females greater performance on extended-response (ER) items (Bolger &

Kellaghan, 1990; Burton, 1996; Lane, Wang, & Magon, 1996; Garner & Engelhard, 1999; Henderson, 1999). Since the removal of the scholarship examinations reduced the number of ER items and increased the importance of the MC items, one would expect a greater proportion of male students to have false positive decisions while a greater proportion of females would have false negative decisions. It appears that for higher achieving students, differences in gender performance due to item format were less pronounced or nonexistent.

The results of this chapter indicate that one in ten students would be incorrectly denied or given a scholarship using the current procedure as compared to the original procedure. Further, it was not possible to attribute these errors to a specific subset of examinations or to gender. Consequently, an attempt was made to reduce the number of errors using alternative psychometric procedures. These results are presented in the following chapter.

CHAPTER 5

ANALYSIS OF THE ALTERNATIVE PROCEDURES

The results of the analyses conducted to address the questions regarding the possibility of improving the decision consistency of the current procedure by incorporating the generalized partial credit model (GPCM), weighting the MC and ER sections, and/or auxiliary information are presented in this chapter. The chapter is organized in three sections. In the first section, the results of the investigation to determine if the GPCM fit the data are presented. Given confirmation that the GPCM could be used, the results for each of the 17 alternative procedures examined are presented and compared to the original, the current, and the GPCM procedures in the second section. The results reported in the first two sections then form the basis to examine psychometric issues in the third section in an attempt to provide possible explanations for some of the findings.

Model/Data Fit for the IRT models used in the Study

Justification for the use of the GPCM and the simultaneous estimation of both the multiple-choice (MC) and extended-response (ER) items was based on the analysis of the fit between the GPCM and the data. Further, as part of the ancillary purpose of the present study, the lack of fit between the model and the data, if it existed, was analysed to determine how it influenced the estimation of scholarship scores. Model data fit was based on the tests of the four assumptions, unidimensionality, item independence, nonspeededness, and lack of guessing, that underlie the use of the unidimensional IRT models.

Unidimensionality. The results for the analysis of unidimensionality are presented in Table 12. The table contains the first five eigenvalues and the proportion of the variance accounted for by the first component obtained from a principal components analysis of the correlation matrix conducted for each of the provincial examinations for 1994/95 and 1995/96, respectively. With the exception of the two Geography examinations, the ratio between the first and second eigenvalues was at least five times and the differences between the other pairs of eigenvalues were close to one for each of the 1994/95 examinations. These findings were confirmed using a scree plot (see 1994/95 Biology, Panel A, Figure 6). Further, the first component in these examinations accounted for at least 15% of the variance and in most cases was close to 20% (see Table 12). Based on these findings it was determined that these examinations met the unidimensionality criteria necessary to justify the use of a unidimensional IRT model.

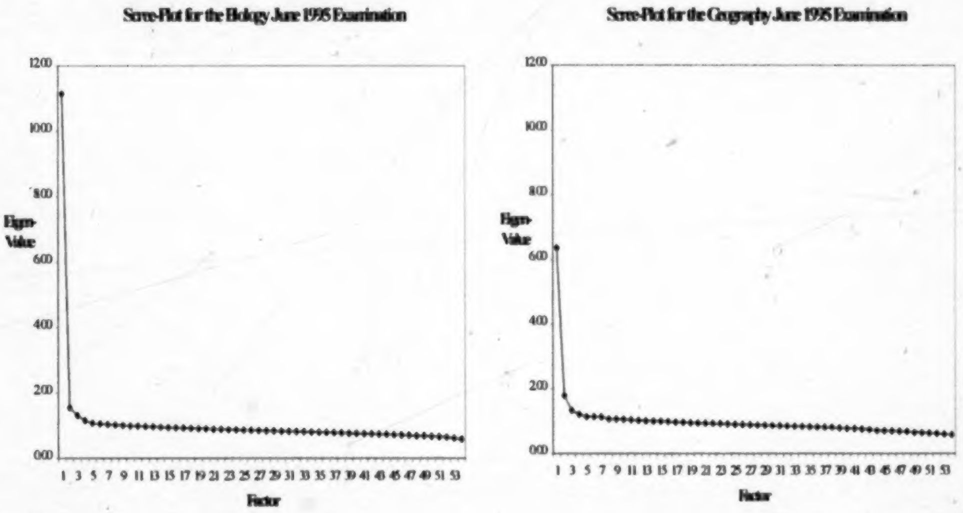


Figure 6. Comparison of Scree Plots for the June 1995 Biology and Geography Examinations

Table 12

The First 5 Eigenvalues (EV) and the Proportion of Variance Accounted for by the First Factor for the Provincial Examinations

	1994/1995						1995/1996					
Subject	EV 1	EV 2	EV 3	EV 4	EV 5	First Factor Variance (%)*	EV 1	EV 2	EV 3	EV 4	EV 5	First Factor Variance (%)*
Biology (January)	11.27	1.41	1.27	1.18	1.13	18.8	11.93	1.63	1.29	1.17	1.11	19.6
Biology (June)	11.14	1.55	1.31	1.16	1.08	18.6	11.60	1.52	1.42	1.16	1.04	19.3
Chemistry (January)	10.36	1.81	1.36	1.19	1.15	17.3	12.13	1.60	1.34	1.21	1.14	20.6
Chemistry (June)	11.49	1.45	1.37	1.08	1.05	19.5	10.96	1.62	1.33	1.14	1.11	18.6
Geography (January)	6.65	1.79	1.40	1.28	1.25	11.8	7.45	1.78	1.59	1.41	1.28	13.5
Geography (June)	6.35	1.80	1.33	1.22	1.14	11.9	7.56	1.62	1.37	1.24	1.18	14.5
Geology (June)	10.25	1.98	1.63	1.44	1.39	13.9	11.35	1.66	1.58	1.40	1.31	15.5
Mathematics (January)	10.22	1.70	1.21	1.17	1.10	17.9	9.93	1.78	1.19	1.15	1.11	17.1
Mathematics (June)	11.55	1.89	1.18	1.05	1.02	20.2	11.27	1.72	1.26	1.12	1.09	19.8
Physics (January)	8.78	1.67	1.25	1.17	1.12	21.9	8.31	1.52	1.28	1.16	1.13	21.3
Physics (June)	8.59	1.61	1.15	1.08	1.05	21.5	8.33	1.48	1.18	1.08	1.06	20.8

* This value represents the proportion of the variance that is accounted for by the first factor expressed as a percentage

In the case of the two Geography examinations the ratio between the first and second eigenvalues was 3.7 and 3.5 times in January and June, respectively. Further, while the differences between the ratios for the other pairs of eigenvalues were close to one, the scree plot suggested a second possible factor (see 1994/95 Geography, Panel B, Figure 6). Lastly, the proportion of variance was below 15%. Based on these results, further analysis of the two Geography examinations was completed. In one analysis, a two-factor model was used to define the examination data. With a two-factor solution, the MC and ER items did separate into two distinct factors in both of the Geography examinations although the ER items tended to load on both factors. In contrast, factor analysis using image with varimax indicated only one factor. Since it was not possible to conclude the two Geography examinations fully met the assumption of unidimensionality, it was determined the GPCM analysis for these examinations would be reviewed more closely to determine if any systematic differences could be found and attributed to problems associated with the apparent lack of unidimensionality.

As summarized in the left panel of Table 12, the 1995/96 examinations better met the unidimensionality criteria required for use of an unidimensional IRT model. The ratio between the first and second eigenvalues was at least five for all of the examinations except Geography. For all of the examinations including Geography the ratios between other pairs of eigenvalues were close to one, indicating that the differences between other eigenvalues were small. The variance accounted for by the first factor was above the 15% minimum in all of the examinations except Geography (slightly below) and was generally close to or above 20%. Further analysis of the two Geography examinations using principal components analysis specifying two factors and image followed by varimax, did not suggest a separate MC and ER component. Based on these results, it was determined that all of the 1995/96 examinations met the criteria to assume essential unidimensionality. Similar results were also obtained when the school-based marks (SBM) were included in the analysis of dimensionality. Thus, for the purposes of using the unidimensional IRT models, the same examinations were also considered unidimensional when SBM was included.

Local item independence. Since all but two of the Geography examinations met the unidimensionality criteria, these same examinations were considered to have met the assumption of local item independence. In the case of the 1994/95 January and June Geography examinations, the assumption of local item independence could not be confirmed or refuted.

Nonspeededness. Based on the response rates of the last three MC items, a minimum of 99% of the students completed all of the MC items on each of the examinations in both 1994/95 and 1995/96. It was determined that all of the MC items were attempted by essentially all of the students. Using the omission rates for the last two ER items, at least 90% of students attempted these items in all of the examinations except Mathematics (all sessions) and Physics (1994/95 only), suggesting that 90% of the students completed the examinations. Comparison of the omission rates for the final two ER items with the omission rates for the previous ER items in the Mathematics and Physics examinations, revealed that the omission rates were similar (10% to 14%) across all ER items. This finding suggests that the omissions had more to do with the difficulty of the ER items rather than lack of time. Thus, all of the examinations were considered to have met the assumption of nonspeededness.

Lack of guessing. The presence of guessing was examined using the *p*-values for both the full sample and the lowest decile of students for each examination. These results are presented in Table 13. The lack of guessing could not be assumed. Some of the *p*-values for the lowest decile of students were larger than that of the entire sample. Further, the *p*-values for the lowest decile of students were generally above 0.15. For this reason, both the two- and three-parameter dichotomous models were considered.

The Use and Effectiveness of the Alternative Procedures

Preliminary analysis comparing the scholarship scores as calculated between the current procedure and the generalized partial credit model (GPCM) procedure revealed that the scholarship scores as calculated by the two procedures did in fact differ. As illustrated previously (see Figure 1, p. 24), a single scholarship score is associated with each provincial examination score using the current procedure. Figure 7 provides the scholarship scores for the 1995 June Chemistry examination if the GPCM procedure was used. With this procedure, different examination scholarship scores are associated with each provincial examination score. This occurs because different items provide differing amounts of information towards the estimation of θ and students with the same raw score would likely have different response patterns. For example, students with the same raw score of 90% had ability estimates expressed on the scholarship scale that varied between 510 and 600 points. In comparison, a provincial examination score of 90% had a single scholarship score of 565 using the current procedure but varied between 485 and 635 using the original procedure (see Figure 1, p. 24).

Table 13

P-values for the three Most Difficult Multiple-Choice Items on Each Examination for the Full sample and the Lowest Decile of Students

	1994/1995			1995/1996		
Subject	Item 1	Item 2	Item 3	Item 1	Item 2	Item 3
Biology (January)	0.28 0.09)	0.34 (0.28)	0.36 (0.14)	0.38 (0.15)	0.42 (0.07)	0.42 (0.21)
Biology (June)	0.42 0.20)	0.42 (0.32)	0.42 (0.16)	0.26 (0.11)	0.29 (0.11)	0.36 (0.13)
Chemistry (January)	0.13 0.23)	0.33 (0.17)	0.33 (0.04)	0.25 (0.15)	0.28 (0.15)	0.46 (0.23)
Chemistry (June)	0.36 0.19)	0.44 (0.17)	0.46 (0.16)	0.29 (0.17)	0.35 (0.18)	0.41 (0.06)
Geography (January)	0.16 0.19)	0.18 (0.14)	0.23 (0.27)	0.17 (0.14)	0.27 (0.12)	0.32 (0.15)
Geography (June)	0.15 0.12)	0.21 (0.10)	0.33 (0.22)	0.21 (0.16)	0.26 (0.13)	0.28 (0.19)
Geology (June)	0.12 0.18)	0.24 (0.16)	0.27 (0.14)	0.23 (0.05)	0.26 (0.25)	0.30 (0.20)
Mathematics (January)	0.26 0.12)	0.35 (0.20)	0.37 (0.20)	0.09 (0.04)	0.36 (0.16)	0.43 (0.15)
Mathematics (June)	0.32 0.15)	0.38 (0.10)	0.42 (0.18)	0.33 (0.09)	0.39 (0.13)	0.39 (0.22)
Physics (January)	0.27 0.10)	0.29 (0.15)	0.31 (0.23)	0.18 (0.19)	0.38 (0.15)	0.41 (0.21)
Physics (June)	0.37 0.29)	0.46 (0.20)	0.48 (0.20)	0.32 (0.12)	0.32 (0.13)	0.49 (0.22)

Note: The *p*-values in brackets are those calculated for the bottom 10% of the students.

Although the correlation between the current and GPCM procedures was 0.98 for the June 1995 Chemistry examination, a comparison of the two figures illustrates that the two procedures did produce different scholarship scores. Further, the distribution of the examination scholarship scores using the GPCM procedure better fit the range of corresponding scores from the original procedure as compared to the current procedure. If the individual scholarship scores using the GPCM procedure were closer to the original scholarship scores than those based on the current procedure, the alternative procedure, in this case the GPCM procedure, could be considered superior.

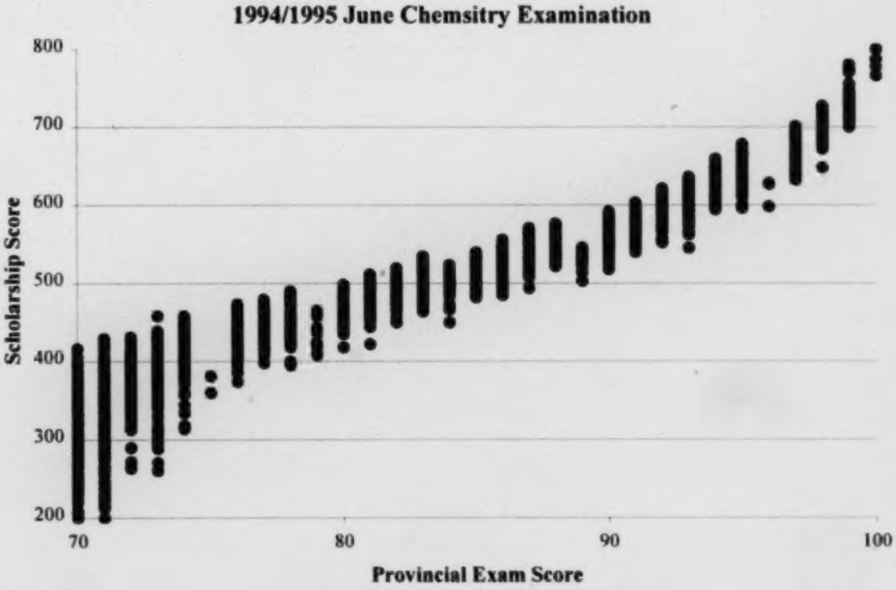


Figure 7. Scholarship Score Distributions for the GPCM Procedure

As described in Chapter 3, three general alternative approaches were investigated in an attempt to reduce the number of errors and the 10% overall error rate that occurred if the current procedure was used in place of the original procedure. The three general procedures incorporated the use of the GPCM, weighting of the multiple-choice (MC) and the extended-response (ER) sections, and/or the use of auxiliary information in the form of school-based mark (SBM). Including modifications to and combinations of these procedures, a total of 17 alternative procedures were considered. Individual examination results, including correlations, root mean square errors (RMSEs), and classification errors for each procedure are presented in Appendix A. Presented in Table 14 are the examination level false negative, false positive, and total error rates for each procedure. The errors are defined by the discrepancy between the decisions made using the procedure in question and the decisions made in 1994/95 and 1995/96 using the original scholarship procedure.

Table 14 is separated into two panels. The left panel contains the results for 1994/95, while the right panel contains the results for 1995/96. The rows in Table 14 are grouped into 4 sections. Section I, containing the results for the current procedure, is included to provide a point of comparison for the alternative procedures. Section II contains the results for the GPCM procedure. The GPCM procedure also serves as a point of comparison for the other procedures that also used the GPCM. In Section III, the results of the six analyses for 1994/95 and four analyses for 1995/96 in which the MC and the ER sections were differentially weighted are presented. Lastly, in Section IV, containing 10 separate analyses for both years, the results for those procedures in which the SBM was used as auxiliary information are provided. As in the previous chapter, both the number of students and the corresponding percentages (in brackets) are reported for both years (see Table 9, p. 78).

Table 14

Error Rates for the Scholarship Decisions using the Alternative Procedures

Alternative Procedure	1994/1995 (N=2524)			1995/1996 (N=2769)		
	False Negative	False Positive	Error Rate (%)	False Negative	False Positive	Error Rate (%)
Section I: Current Procedure	164 (6.5)	85 (3.4)	9.9	195 (7.0)	89 (3.2)	10.3
Section II: GPCM	167 (6.6)	82 (3.2)	9.9	201 (7.3)	104 (3.8)	11.0
Section III: Weighting						
Classical, (Scholarship)	172 (6.8)	81 (3.2)	10.0	212 (7.7)	92 (3.3)	11.0
Classical, (Optimal)	172 (6.8)	83 (3.3)	10.1	194 (7.0)	90 (3.3)	10.3
GPCM, (2-parm; Scholarship)	175 (6.9)	83 (3.3)	10.2	222 (8.0)	103 (3.7)	11.7
GPCM, (2-parm; Optimal)	177 (7.0)	76 (3.0)	10.0	209 (7.5)	100 (3.6)	11.2
GPCM, (3-parm; Scholarship)	180 (7.1)	82 (3.2)	10.4	- -	- -	-
GPCM, (3-parm; Optimal)	181 (7.2)	80 (3.2)	10.3	- -	- -	-
Section IV: Auxiliary Information						
Classical, MC-ER-SBM	153 (6.1)	95 (3.8)	9.8	187 (6.8)	99 (3.6)	10.3
Classical, SBM (Optimal)	160 (6.3)	95 (3.8)	10.1	187 (6.8)	95 (3.4)	10.2
Classical; MC/ER/SBM (Optimal)	155 (6.1)	90 (3.6)	9.7	185 (6.7)	89 (3.2)	9.9
GPCM, MC-ER-SBM	153 (6.1)	92 (3.6)	9.7	192 (6.9)	111 (4.0)	10.9
GPCM, MC-ER/SBM (0.9, 0.1)	159 (6.3)	82 (3.2)	9.5	199 (7.2)	105 (3.8)	11.0
GPCM, MC-ER/SBM (Optimal)	158 (6.3)	83 (3.3)	9.5	199 (7.2)	109 (3.9)	11.1
GPCM, MC/ER/SBM (0.45,0.45, 0.1)	161 (6.4)	86 (3.4)	9.8	193 (7.0)	106 (3.8)	10.8
GPCM, MC/ER/SBM (Optimal)	162 (6.4)	84 (3.3)	9.7	199 (7.2)	99 (3.6)	10.8
GPCM, MC/ER-SBM (0.5, 0.5)	147 (5.8)	92 (3.6)	9.5	189 (6.8)	111 (4.0)	10.8
GPCM, MC/ER-SBM (Optimal)	153 (6.1)	89 (3.5)	9.6	196 (7.1)	106 (3.8)	10.9

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of students in each category.

Scholarship Decision Consistency Using the GPCM

As shown in the left panel (1994/95) of Section II on Table 14, 167 (6.6%) students who were given a scholarship using the original procedure would not have been given a scholarship using the GPCM (false negative). Further, 82 (3.2%) students who were not given a scholarship using the original procedure would have been given a scholarship based on the GPCM (false positive). Combining these two error rates, the total error rate was 9.9%. This was the same overall error rate as found for the current procedure (see Section 1, Table 14). For 1995/96 (see the right panel) GPCM results, 201 (7.3%) students incorrectly would have been denied a scholarship and 104 (3.8%) students incorrectly would have received a scholarship for a total error rate of 11.0%. The total error rate was slightly higher than the 10.3% total error rate found for the current procedure. This difference likely reflects random variation across the years since the correlations, RMSEs, and classification errors for the current and GPCM procedures were generally comparable across both years (see Appendix A, Table 19). Consequently, using the GPCM alone did not improve upon the errors associated with the current procedure and can not be justified as an alternative psychometric method to increase decision consistency in the present context.

Scholarship Decision Consistency Using Subtest Weighting

As described in Chapter 3 (see p. 63), two different weighting approaches were used: scholarship weights with the MC and ER sections weighted to have the same contribution to the total score as in the original procedure (scholarship) and optimal weights with regression weights used to define the contribution of each section (optimal). These weighting approaches were used with the current procedure (classical) and the

GPCM using both the two-parameter model (GPCM (2-parm)) and the three-parameter model for the dichotomously-scored items (GPCM (3-parm)). Due to difficulties achieving convergence when the three-parameter model was specified in PARSCALE 3.1 and given that the three-parameter model did not produce better results than the two-parameter model, the approaches using the three-parameter model were not completed for the 1995/96 examinations.

For 1994/95, the use of weighting, regardless of the combination used, did not improve upon the results obtained with the current or GPCM procedures (see section 3, Table 14). Again, approximately one in 10 students would have been misclassified using any of these six procedures. However, in comparison to the current or GPCM procedures alone, the use of weighting did yield a slightly greater number of false negative decisions and a slightly smaller number of false positive decisions with the change in false negative decisions being slightly larger.

As with the 1994/95 examinations, the use of weighting in the 1995/96 examinations did not improve upon the results obtained using the current procedure. With the exception of the classical model using optimal weighting (Classical, (Optimal)), the use of weighting seemed to slightly increase the number of false negative decisions but had a random effect on the false positive decisions. Based on the two years analysed, the use of weighting can not be justified as a psychometric method to increase decision consistency beyond that of the current procedure. Further, while the false negative rates did marginally increase in most instances, it is not possible to determine if these differences were due to random variation or systematic effects associated with the use of weighting.

Scholarship Decision Consistency using School-Based Mark as Auxiliary Information

The results for the 10 different procedures that incorporated school-based mark (SBM) are reported in section IV of Table 14. As described in chapter 3 (see p. 65), the procedures differed first in the use of either the classical or GPCM approach and second in the manner in which SBM was combined with the provincial examinations. For the 1994/95 school year, the use of SBM provided marginal improvement over the current and GPCM procedures alone in most instances. While similar results were found for the classical combinations using SBM for 1995/96, the GPCM combinations using SBM were unable to equal the error rate of the current procedure. Taken together, the results across the two years failed to support the use of the approaches that included SBM. Again, the total error rate was approximately 10%. However, in contrast to subtest weighting of the MC and ER sections alone, the use of SBM generally reduced the false negative rate and, to a lesser extent, increased the false positive rate as compared to the current or GPCM procedures. The effect was consistent across both years although it was more apparent in 1994/95. Nonetheless, given the marginal differences, it remains to be shown if these variations are related to the use of SBM or are due to random variation.

Summary

Given the overall similarity between the error rates across the different approaches, none of the alternative approaches were found to be superior to the current procedure to the extent it could be recommended as an approach that would provide better decision consistency. The results at the examination level (summarized in Appendix A) also failed to indicate that one approach could be considered superior. Due to the similarity in the results across all of the approaches in comparison to the current procedure at both the individual examination and the total scholarship score levels, it was determined that further investigation exploring the alternative procedures for subsets of examinations would provide little further insight. However, the potential interaction between the alternative procedures and false negative and false positive error rates, although small and difficult to distinguish from random error, becomes important from a psychometric perspective if it supports or refutes previous research regarding similar issues. Before examining the psychometric perspectives as related to the alternative procedures, a final set of analyses was completed examining the decision consistency for both males and females using each alternative procedure.

The Effect of Gender on Decision Consistency for the Alternative Procedures

The results for males and females are summarized in Table 15 for 1994/95 and in Table 16 for 1995/96. The table are presented using the same format as used in Table 14 (see p. 98) with the exception that the left panel contains the results for females and the right panel the results for males. As with the current procedure, the percentage of total errors was greater for males than females regardless of the procedure used or the year analysed. There were some marginal variations in the false negative and false positive rates for both males and females depending on the general approach used, for example, weighting or the inclusion of SBM, but the results did not fit any expected result or follow any consistent pattern.

Table 15

Error Rates for the Scholarship Decisions Based on Gender using the Alternative Procedures for 1994/95

	Female (N=1088)			Male (N=1436)		
Alternative Procedure	False Negative	False Positive	Error Rate (%)	False Negative	False Positive	Error Rate (%)
Section I: Current Procedure	57 (5.2)	29 (2.7)	7.9	107 (7.5)	56 (3.9)	11.4
Section II: GPCM	57 (5.2)	29 (2.7)	7.9	110 (7.7)	53 (3.7)	11.4
Section III: Weighting						
Classical, (Scholarship)	60 (5.5)	30 (2.8)	8.3	112 (7.8)	51 (3.6)	11.4
Classical, (Optimal)	61 (5.6)	28 (2.6)	8.2	111 (7.7)	56 (3.9)	11.6
GPCM, (2-parm; Scholarship)	58 (5.3)	33 (3.0)	8.4	117 (8.1)	50 (3.5)	11.6
GPCM, (2-parm; Optimal)	59 (5.4)	28 (2.6)	8.0	118 (8.2)	48 (3.3)	11.6
GPCM, (3-parm; Scholarship)	56 (5.1)	32 (2.9)	8.1	124 (8.6)	50 (3.5)	12.1
GPCM, (3-parm; Optimal)	65 (6.0)	33 (3.0)	9.0	116 (8.1)	47 (3.3)	11.4
Section IV: Auxiliary Information						
Classical, MC-ER-SBM	47 (4.3)	36 (3.3)	7.6	106 (7.4)	59 (4.1)	11.5
Classical, SBM (Optimal)	51 (4.7)	36 (3.3)	8.0	109 (7.6)	59 (4.1)	11.7
Classical; MC/ER/SBM (Optimal)	50 (4.6)	34 (3.1)	7.7	105 (7.3)	56 (3.9)	11.2
GPCM, MC-ER-SBM	43 (4.0)	42 (3.9)	7.8	110 (7.7)	50 (3.5)	11.1
GPCM, MC-ER/SBM (0.9, 0.1)	50 (4.6)	31 (2.8)	7.4	109 (7.6)	51 (3.6)	11.1
GPCM, MC-ER/SBM (Optimal)	47 (4.3)	32 (2.9)	7.3	111 (7.7)	51 (3.6)	11.3
GPCM, MC/ER/SBM (0.45,0.45, 0.1)	46 (4.2)	34 (3.1)	7.4	115 (8.0)	52 (3.6)	11.6
GPCM, MC/ER/SBM (Optimal)	48 (4.4)	34 (3.1)	7.5	114 (7.9)	50 (3.5)	11.4
GPCM, MC/ER-SBM (0.5, 0.5)	41 (3.8)	42 (3.9)	7.6	106 (7.4)	50 (3.5)	10.9
GPCM, MC/ER-SBM (Optimal)	46 (4.2)	42 (3.9)	8.1	107 (7.5)	47 (3.3)	10.7

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of students in each category

Table 16

Error Rates for the Scholarship Decisions Based on Gender using the Alternative Procedures for 1995/96

Alternative Procedure	Female (N=1226) False Negative	Female (N=1226) False Positive	Female (N=1226) Error Rate (%)	Male (N=1543) False Negative	Male (N=1543) False Positive	Male (N=1543) Error Rate (%)
Section I: Current Procedure	78 (6.4)	33 (2.7)	9.1	117 (7.6)	56 (3.6)	11.2
Section II: GPCM	84 (6.9)	40 (3.3)	10.1	117 (7.6)	64 (4.1)	11.7
Section III: Weighting						
Classical, (Scholarship)	87 (7.1)	34 (2.8)	9.9	125 (8.1)	58 (3.8)	11.9
Classical, (Optimal)	80 (6.5)	32 (2.6)	9.1	114 (7.4)	58 (3.8)	11.1
GPCM, (2-parm; Scholarship)	95 (7.7)	42 (3.4)	11.2	127 (8.2)	61 (4.0)	12.2
GPCM, (2-parm; Optimal)	90 (7.3)	38 (3.1)	10.4	119 (7.7)	62 (4.0)	11.7
GPCM, (3-parm; Scholarship)	- -	- -	-	- -	- -	-
GPCM, (3-parm; Optimal)	- -	- -	-	- -	- -	-
Section IV: Auxiliary Information						
Classical, MC-ER-SBM	77 (6.3)	40 (3.3)	9.5	110 (7.1)	59 (3.8)	11.0
Classical, SBM (Optimal)	78 (6.4)	37 (3.0)	9.4	109 (7.1)	58 (3.8)	10.8
Classical; MC/ER/SBM (Optimal)	77 (6.3)	33 (2.7)	9.0	108 (7.0)	56 (3.6)	10.6
GPCM, MC-ER-SBM	78 (6.4)	47 (3.8)	10.2	114 (7.4)	64 (4.1)	11.5
GPCM, MC-ER/SBM (0.9, 0.1)	84 (6.9)	43 (3.5)	10.4	115 (7.5)	62 (4.0)	11.5
GPCM, MC-ER/SBM (Optimal)	84 (6.9)	45 (3.7)	10.5	115 (7.5)	64 (4.1)	11.6
GPCM, MC/ER/SBM (0.45,0.45, 0.1)	83 (6.8)	44 (3.6)	10.4	110 (7.1)	62 (4.0)	11.1
GPCM, MC/ER/SBM (Optimal)	86 (7.0)	39 (3.2)	10.2	113 (7.3)	60 (3.9)	11.2
GPCM, MC/ER-SBM (0.5, 0.5)	83 (6.8)	48 (3.9)	10.7	106 (6.9)	63 (4.1)	11.0
GPCM, MC/ER-SBM (Optimal)	85 (6.9)	44 (3.6)	10.5	111 (7.2)	62 (4.0)	11.2

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of students in each category

Given the research discussed earlier that MC items tend to favour males while ER items tend to favour females and since MC items have implicitly more weight in the current procedure (the scholarship examinations only contained ER items), the increased weighting of the ER section was expected to lower the false negative rate for females and the false positive rate for males in comparison to the current procedure (e.g., Bolger & Kellaghan, 1990; Garner & Engelhard, 1999; Henderson, 1999). However, the false negative rate increased for both genders and the false positive rate only decreased marginally for males in 1994/95. Thus, the use of increased weighting of the ER section (scholarship weight) did not differentially effect either gender.

Similarly, the use of SBM, another factor that could be expected to differentially effect the two genders, had different effects across the two years. In 1994/95, the use of SBM tended to lower the false negative rate and to a lesser extent, increase the false positive rate for females in comparison to the current procedure. In 1995/96, the false negative rate tended to remain constant and again the false positive rate slightly increased for females. Further, the false negative rate, which tended to be similar to the current procedure during the 1994/95 year, was marginally smaller than the rate for the current procedure during the 1995/96 year. The error rates for males were not affected either year. As with the use of weighting, the use of SBM did not consistently nor differentially effect the decision error rates for either males or females as compared to the current procedure suggesting that the differences were due to random variation rather than systematic variations attributable to gender. As reported for the current procedure, there were no differences in the results of the alternative procedures that could be attributed to the gender of the higher achieving students who were part of the present study.

Psychometric Issues Associated with the Alternative Procedures

The purpose of this section is to examine four psychometric issues that arose in the use of the alternative procedures. These issues were: 1) the effects of model data fit on ability estimation; 2) the differences between simultaneous estimation, separate estimation, and weighting of the MC and ER sections; 3) the effect of SBM on scholarship score accuracy; and 4) the stability of the 3-parameter model for dichotomously-scored items as used in PARSCALE 3.1.

The Effects of Model Data Fit on Ability Estimation

Given that the assumptions required to conclude adequate model data fit were not equally met by all of the examinations, the first psychometric issue that arose was the effect of weaker model data fit on the accuracy of the ability estimates. In particular, the assumption of unidimensionality could not be fully justified for the two 1994/95 Geography examinations. These two examinations also had the poorest results for that year, in terms of correlations, RMSEs, and classification error rates (see, Appendix A, Table 19). However, the results for the Geography examinations using the GPCM were comparable to the correlations, RMSEs, and classification error rates using the current procedure (see Table 7, p. 74 and Table 9, p. 78). If poor results in the GPCM procedure can be linked to poor model/data fit, poor fit may also be problematic within the classical framework. On the other hand, poor correlations, RMSEs, and classification errors were also obtained for the 1995/96 Geography examinations even though these examinations better met the criteria used to establish essential unidimensionality as compared to the 1994/95 Geography examinations.

The other assumption that was not fully met was the lack of guessing. The results of this analysis prompted the use of the 3-parameter model for the dichotomously-scored examination items. However, the three-parameter model often failed to converge and when it did converge, yielded results with lower correlations and higher RMSEs than those obtained using the 2-parameter model. Due to these problems, the 3-parameter model was abandoned. However, it became a psychometric issue to be analysed further. Based on these findings, in spite of observed differences in how well the data from the different examinations met the assumptions, the ability estimates were largely unaffected.

The Differences Between Simultaneous Estimation, Separate Estimation, and Weighting of the MC and ER Sections

Previous research (e.g., Wilson & Wang, 1995) has suggested that ER items provide more information for higher ability students than dichotomously-scored items. Further, the separation of MC and ER items has been proposed when IRT models are being used because of the different cognitive dimensions measured by each format (Luecht, 1994). These findings prompted the use of analytical procedures that separated and weighted the MC and ER sections of the provincial examinations (see section III, Table 14). However, the results did not justify the separation of the MC and the ER sections nor did they support the use of increased weighting of the ER sections on the individual examinations. Similar results were obtained regardless of whether the estimates were obtained based on simultaneous or separate estimation of the two sections. The results in Appendix A (see Table 20 to Table 25) and Table 14 show that there was little consistent difference in either the scholarship examination scores or the decision consistencies in procedures using either simultaneous or separate estimation.

Additional evidence was found in the examination of the item parameters. The June 1995 Geography and Chemistry examinations were used to compare the similarity in item parameters for three different estimation methods: simultaneous estimation of the MC and ER sections (simultaneous), separate estimation of the MC and the ER sections (separate), and simultaneous estimation of the MC, ER, and SBM (combined). These two examinations provided some contrast not only because of the differing focus of the examinations but also because of the generally poorer results obtained for the June 1995 Geography examination and the generally superior results for the June 1995 Chemistry examination in terms of meeting the assumptions and the quality of the results. Table 17 contains the means and standard deviations for the item parameters for both the MC and ER sections based on each of the three estimation methods. Table 18 contains the correlations between the *a*- and *b*-parameters for the MC section and the *a*- and the first four *b*-parameters for the ER section among these three estimation methods.

If the MC and ER sections were measuring different latent traits, it would be expected that the item parameters would vary depending on whether the two sections were estimated simultaneously or separately. In fact, variations in the item parameters across different methods could be used as further evidence of multidimensionality. The results show that the item parameters had similar means and standard deviations and were closely correlated across the three methods for Chemistry (see left panel). Marginal differences did exist in the item parameters when the MC and the ER sections were estimated separately in comparison to the two simultaneous procedures for Geography (see right panel), perhaps providing additional evidence that this examination may not have fully met the assumptions of unidimensionality.

Table 17

Descriptive Statistics for Item Parameters for Different Estimation Procedures

	June 1995 Chemistry						June 1995 Geography					
	MC/ER Simultaneous		MC-ER Separate		MC/ER/SBM Combined		MC/ER Simultaneous		MC-ER Separate		MC/ER/SBM Combined	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Multiple Choice												
A-Parameter	0.67	0.19	0.67	0.19	0.67	0.19	0.32	0.13	0.35	0.13	0.32	0.12
B-Parameter	-0.99	0.85	-0.99	0.86	-0.98	0.85	-0.20	1.73	-0.25	1.63	-0.19	1.72
Extended Response												
A-Parameter	0.29	0.09	0.30	0.10	0.29	0.09	0.21	0.07	0.22	0.09	0.21	0.07
B1	0.42	2.27	0.46	2.39	0.42	2.25	0.09	2.86	0.21	2.79	0.10	2.85
B2	-1.73	3.94	-2.00	4.02	-1.72	3.91	-2.24	5.56	-2.43	5.00	-2.23	5.53
B3	1.09	2.43	1.11	2.50	1.08	2.42	-2.42	5.59	-2.05	4.86	-2.43	5.56
B4	-2.15	2.72	-2.15	2.86	-2.14	2.69	-1.29	4.93	-1.44	4.39	-1.30	4.87

Table 18

Item Parameter Correlations between Different Estimation Procedures

	June 1995 Chemistry			June 1995 Geography		
	Simult. / Separate	Simult. / Combined	Separate / Combined	Simult. / Separate	Simult. / Combined	Separate / Combined
Multiple Choice						
A- Parameter	1.00	1.00	1.00	0.98	1.00	0.98
B- Parameter	1.00	1.00	1.00	1.00	1.00	1.00
Extended Response						
A-Parameter	0.97	1.00	0.97	0.97	1.00	0.98
*B*1	1.00	1.00	1.00	1.00	1.00	1.00
*B*2	1.00	1.00	1.00	1.00	1.00	1.00
*B*3	0.99	1.00	0.99	1.00	1.00	1.00
*B*4	1.00	1.00	1.00	1.00	1.00	1.00

Nonetheless, a comparison of the correlations, RMSEs, and classification consistencies across these estimation methods (see Appendix A, Table 19, Table 23, and Table 28) for the June 1995 Geography examination did not indicate there were differences attributable to simultaneous or separate estimation. Further analysis of the other examinations could be used to determine if the minor differences found in Geography represented random variation or were due to an interaction between highly correlated but different latent traits being measured by the MC and ER sections. Such analysis could also be used to determine the use of such an approach to test for unidimensionality.

As with the separation of the MC and ER items, the differential weighting of the MC and ER sections had a minimal and inconsistent effect on the error rates. Given that research has shown that ER items better differentiate higher ability students one would expect the false negative decisions to decrease when the ER sections were weighted more heavily (e.g., Wilson & Wang, 1995; Carlson, 1999). To the extent that the provincial examination ER items were more closely related to the scholarship examination ER items than between the provincial examination MC items and the scholarship examination ER items, it was expected that the ER section of the provincial examinations would be a better predictor of scholarship scores. However, the ER section was rarely the best predictor of scholarship scores, nor did the beta weights consistently increase the weighting of this section in comparison to the MC section for those procedures using optimal weighting. Even the weighting approach that purposely increased the weight of the ER section (scholarship) did not improve upon the false negative decisions. If anything, the false negative errors increased when the weight of the ER section was increased beyond its natural weight in the total score metric.

The Use of School-Based Mark as Auxiliary Information

The use of a student's school-based mark (SBM) provided an opportunity to examine the accuracy of scholarship scores and decisions when other information could be used alongside the provincial examinations. From a policy perspective, the advantage of using SBM is that there is no additional cost to the examination process since the SBM is available for all students. From a psychometric perspective, the current study focused on the best way to incorporate SBM as auxiliary information, as an additional test item or as a separate subtest. Although the overall error rates were very similar, the use of SBM did seem to marginally reduce the false negative rates while generally increasing the false positive rates. The reduction in false negative decisions could be due to the SBM compensating those students who had a relatively weaker performance on a provincial examination in comparison to their performance on the corresponding scholarship examination. Further, as indicated by the increase in false positive rate, the SBM may have also compensated those students who had a relatively weaker performance on a scholarship examination in comparison to their performance on the corresponding provincial examination. While the effects were small, the SBM may have had the effect of offsetting some of the differences in individual student performance across the provincial and scholarship examinations in the original procedure.

Finally, although the IRT item parameter estimates for the MC and ER items were similar regardless of whether the MC, ER, and SBM were estimated simultaneously or separately, the parameters for the SBM varied. For example, for the January 1994/95 Biology examination, the *a*-parameter (discrimination) for SBM was 0.90 when it was treated as a single item subtest; however, when SBM was included with the MC and the ER items, its *a*-parameter was 1.19. Similarly, for the June 1995 Chemistry examination, the *a*-parameter increased from 0.50 to 1.08 when SBM was estimated along with the MC and ER items rather than alone. These findings would support the supposition by Yen (1986) that item discrimination increases when multidimensional items are estimated using an unidimensional IRT model. While this trend was found in other examinations, the opposite result was also found. For example, in the June 1995 Geography examination, the *a*-parameter dropped from 1.00 to 0.60 when SBM was estimated simultaneously with the other items, a finding in line with the research of De Ayala (1995). Since both increases and decreases were observed in the value of the *a*-parameter for SBM when it was estimated with the MC and ER items, it was not possible to provide definitive support for the findings of either Yen (1986) or De Ayala (1995). Nonetheless, the change in the value of the *a*-parameter itself, depending on the way the items were combined in order to be estimated, could be considered evidence of multidimensionality. The results suggest that SBM was measuring a related but different trait than the provincial examination items. However, since the SBM was a single item, it seemed to have little effect on the estimates of either the MC or the ER sections of the examination (see Table 17) or the estimates of ability (see Appendix A).

The 3-Parameter Model in PARSCALE 3.1

As pointed out earlier, the 3-parameter model failed to converge when used with many of the examinations. In using the 3-parameter model with the dichotomously-scored items alone in PARSCALE 3.1, the program would often fail to reach the convergence criterion even after 200 cycles and in some examinations, the process would stop unexpectedly due to extremely large changes within the EM cycle. Further, the item parameter estimates were often quite extreme with the *a*-parameter above 2.0 and the *b*-parameter either below –4.0 or above 4.0. These results were unexpected. Therefore, the June 1995 Chemistry and Geography examinations were reanalysed using the 3-parameter model within BILOG 3.11 (Mislevy & Bock, 1990). These two examinations were used because convergence and reasonable estimates were obtained with the 3-parameter model in PARSCALE 3.1 for the Chemistry items but not for the Geography items.

For both examinations, the estimation process in BILOG 3.11 under default conditions was able to converge and yielded reasonable item parameter estimates for all of the items. In the case of Chemistry, the results yielded by both PARSCALE 3.1 and BILOG 3.11 were similar. The ability (θ) estimates derived from PARSCALE 3.1 had a mean of –0.07 and a standard deviation of 0.96, while the corresponding values derived from BILOG 3.11 were 0.02 and 1.17. The correlation between the two distributions of θ estimates was 0.99. With respect to item parameters, using PARSCALE 3.1, the *a*-, *b*-, and *c*-parameters had a mean of 0.83, -0.63, and 0.23, respectively, with corresponding standard deviations of 0.25, 1.05, and 0.14. The corresponding means obtained using BILOG 3.11 were 0.81, -0.52, and 0.23 and the corresponding standard deviations were 0.22, 0.96, and 0.08. The item parameter correlations between the two programs were 0.99 for the *a*-parameter, 0.91 for the *b*-parameter, and 0.88 for the *c*-parameter. Although the correlations for the *b*- and *c*-parameters were substantially lower than that of the *a*-parameter, it was concluded that similar results would be obtained for the June 1995 Chemistry examination regardless of the computer program used.

In the case of Geography, the results were less similar. The ability estimates derived from PARSCALE 3.1 had a mean of –0.56 and a standard deviation of 1.18, while the corresponding values derived from BILOG 3.11 were -0.03 and 1.24. The correlation in the θ estimates between the two programs was 0.91. With respect to the item parameters, using PARSCALE 3.1, the *a*-, *b*-, and *c*-parameters had a mean of 0.75, 0.56, and 0.30, respectively, with corresponding standard deviations of 0.21, 0.78, and 0.76. The corresponding means obtained using BILOG 3.11, were 0.61, 0.62, and 0.26 while the corresponding standard deviations were 0.21, 0.78, and 0.76. The item parameter correlations between the two programs were 0.21 for the *a*-parameter, 0.78 for the *b*-parameter, and 0.76 for the *c*-parameter. Based on these results, it was concluded that the use of PARSCALE 3.1 produced different results than what would have been obtained if BILOG 3.11 was used for the June 1995 Geography examination.

To the extent that the BILOG 3.11 results can be considered to produce results that are closer to the actual parameters, this exploratory analysis indicates there were some problems with the use of the 3-parameter model in PARSCALE 3.1. These findings may explain the generally weaker results and poorer decision consistencies obtained when the 3-parameter model was used in the present study. However, it is not known how much of an effect better estimation with the 3-parameter model would have had on the overall findings. Given that the other alternative procedures failed to improve upon the current procedure it is expected any improvements would have been marginal at best. Nonetheless, the problems found in this secondary analysis do require further study in order to clarify expectations and identify potential estimation problems. For example, if convergence is not obtained within PARSCALE 3.1, one can expect the results to be quite different from those obtained using programs designed solely for dichotomously-scored items. In the case of the June 1995 Geography examination, the extreme item parameters obtained in PARSCALE 3.1 were not obtained when BILOG 3.11 was used. This would explain the lower correlations found in the Geography example. Further, the parameter distributions seem to differ somewhat when the 3-parameter model was used in PARSCALE 3.1. For example, the standard deviations for the item parameters were higher in PARSCALE 3.1 as compared to BILOG 3.11 but comparatively lower for the ability estimates.

CHAPTER 6

SUMMARY AND CONCLUSIONS

The research questions and a brief description of the methods used in this study are presented first in this chapter followed by a summary of the key findings. The limitations of the study are then presented. The conclusions and implications for practice from both a policy and measurement perspective are then discussed in light of the limitations. The chapter concludes with a series of recommendations for future research.

Summary of Research Questions and Methods

In 1996/97, the British Columbia Ministry of Education modified its scholarship examination process by removing the optional scholarship examinations, the scores from which were combined with the scores yielded by the required provincial examinations to determine scholarship recipients from the population of graduating High-School students. In removing the scholarship examinations, the Ministry adopted the current procedure that uses only the provincial examination scores to calculate examination scholarship scores and identify scholarship recipients. To the extent that the reduction in the length and difficulty of the examinations used to determine scholarship scores has caused differences in the identification of scholarship recipients, it is important to determine if such differences can be reduced through alternative cost-effective approaches. Consequently, the primary purposes of this study were to determine 1) the changes in the scholarship decisions that occurred due to the change in the scholarship procedure being used by the British Columbia Ministry of Education and 2) if there is an alternative procedure that could be used to reduce any of the random or systematic errors in the scholarship decisions that have occurred due to the change in the scholarship procedure.

More specifically the following two primary questions and one ancillary question were addressed in this study:

1. How has the elimination of the one-hour scholarship examinations changed which students receive scholarships?
2. Can the use of alternative approaches incorporating item response theory in the form of the Generalized Partial Credit Model (GPCM), weighting of the MC and ER sections, and/or including auxiliary information in the form of school-based mark improve upon the decisions made?
3. Do theorized interactions and differences occur when alternative approaches are compared?

In order to assess the impact of the change in the scholarship procedure and the potential of the alternative approaches, the student data from a sample of 11 provincial examinations was used. Replication of the results was accomplished through the use of two consecutive years of data, 1994/95 and 1995/96. Along with an examination of the current procedure, a total of 17 alternative procedures were investigated in an attempt to reduce the differences between the original and current procedures. The alternative procedures included, individually or in combination, IRT methods of score estimation, weighting of the multiple-choice and extended-response sections of the examinations, and the use of auxiliary information in the form of school-based mark.

Using the results from the original scholarship procedure as the standard for comparison, the analysis of the differences between the original and alternative procedures were reported using correlations, root mean square errors (RMSEs), and decision consistencies at the examination and total scholarship score levels. With respect to the original procedure, increased correlations, reduced RMSEs, and lower decision consistencies at the examination and total scholarship score levels were used as evidence of a superior approach.

Findings

Comparison of the Original and Current Procedures

The use of the current procedure affected the scholarship scores at the examination and total scholarship score levels. At the examination level, correlations between the original and the current procedures ranged from a low of 0.71 to a high of 0.92 with a median correlation of 0.88 across the two years considered. Similarly, the RMSEs varied from 38.68 to 74.03 with the mean RMSE being 53.97 or 9.0% of the score range. In comparison to the original procedure, the classification error rates varied from 8.1% to 22.7% with an average of 11.4% across the examinations. In particular, the false negative rates, those scholarship scores that were below 475 using the current procedure but above 475 in the original procedure, were much larger than the false positive rates, being on average over five times larger.

The use of the sum of each student's three highest scholarship scores above 475 to determine the total scholarship score did mediate some of the errors associated with the individual examinations but the error rate remained substantial. The classification error rate for the awarding of scholarships based on the total scholarship score was 9.9% in 1994/95 and 10.3% in 1995/96 indicating that approximately one in 10 students would be incorrectly denied or given a scholarship using the current procedure. Further, the false negative error rate was approximately two times the false positive rate. Hence, the majority of the decision errors were such that students would be unfairly denied a scholarship using the current procedure. An examination of the influence of individual examinations or combinations of examinations failed to provide substantive evidence that the overall error rates could be attributed to any specific examination or combination of examinations. However, there were indications that a higher preponderance of false negative decisions occurred in association with what are considered the more difficult academic courses (e.g., Chemistry, Physics, and Mathematics).

Lastly, a comparison of the error rates for males and females indicated that fewer females than males had a minimum of three examination scholarship scores necessary to receive a total scholarship score based on the examinations considered in this study. In 1994/95, 1,088 females versus 1,436 males had a minimum of three examination scholarship scores and in 1995/96 the numbers were 1,226 and 1,543, respectively. Further, the proportion of decision errors for females was lower than that for males, 7.8% versus 11.4% in 1994/95 and 9.1% versus 11.2% in 1995/96. However, the ratio of false negative to false positive decisions was virtually the same for both genders at approximately two to one.

Comparison of the Alternative Procedures to the Original and Current Procedures

While the scholarship scores calculated for each examination using the current procedure differed from those calculated using the alternative procedures, the differences were such that the correlations and RMSEs between the original and the alternative procedures were virtually the same as those between the original and the current procedures. Further, the classification and decision error rates at both the examination and the total scholarship score levels of the current procedure were comparable to those of the alternative procedures. In terms of the scholarship decisions at the total scholarship score level, the alternative procedures did not reduce the approximately one in 10 error rate associated with the current procedure nor did these alternative procedures lower the proportionally higher false negative rate. Due to the similarity between the current and the alternative procedures, further analyses using specific examinations or combinations of examinations were not completed.

The 17 alternatives procedures were applied separately for both males and females and the results compared. Again, the proportion of total errors was greater for males than females across the 17 alternative procedures and none of the procedures was found to differentially affect males and females in terms of the total error rate. Marginal variations in the false negative and false positive error rates for males and females were observed for those procedures using weighting or SBM; however, the variations were not systematic.

Psychometric Issues

In examining and comparing the results of the current and the 17 alternative procedures to those obtained using the original procedure, four psychometric issues were addressed that were pertinent to the results of the study as well as to previous research. The findings for these issues were as follows:

1. The link between weaker model data fit and poorer estimation of ability within the GPCM was either not realized or was found to be comparable in both the GPCM and the classical total test score models.
2. There were no substantial differences in either the scores or scholarship decisions obtained when the MC and ER items were simultaneously estimated, separately estimated, or differentially weighted.

3. The school-based mark (SBM) did not have an impact on the overall decision consistency but marginally decreased the proportion of false negative decisions and marginally increased the proportion of false positive decisions. Second, within the context of IRT, the item parameter estimates for SBM varied depending on the estimation method used.
4. The three-parameter model as used in PARSCALE 3.1 often produced poor results with unreasonable parameter estimates. Further, these results differed from those that were obtained using BILOG 3.11.

Limitations

In considering the scholarship examination program in British Columbia, the effects of the current procedure and the possible use of alternative procedures to better determine scholarship recipients was previously delimited to a subset of 11 of the 25 provincial examinations. Six examinations could not be included because of low sample size (German, Japanese, Latin, Mandarin, and Spanish). Two examinations were not included because one form of the examination was compromised (January and June French 12). Five examinations were excluded because of the use of holistic scoring scales (English Literature, Français Langue, and January and June History). Finally, the January and June English 12 were not included because these examinations did not have a scholarship examination.

In completing the analyses based on these examinations the study was further limited in four ways. First, the current study reported findings based on two years of examinations, 1994/95 and 1995/96. It is possible that the results from the two years investigated were unique in their relation between the provincial and scholarship examinations. Of more concern are the small differences observed between the current and alternative procedures across the two years. Since only two years were included in the study, the small differences that were observed could not be attributed to systematic differences or random variation amongst the procedures. Nonetheless, even if the marginal differences amongst the alternative procedures were found to be consistent over a greater time period, the results indicate the effects are small.

Second, while the number of students who completed at least three scholarship examinations and also achieved at least 70% on three provincial examinations was large, the number of possible examination combinations in which the scholarship scores could be based was also large and thus, the number of students with any specific combination was generally small. The small number of students with any specific combination of examinations made it difficult to examine the differential impact of various examination combinations of the 11 examinations included in the present study.

Third, in order to examine aspects of differential weighting of the multiple-choice (MC) and the extended-response (ER) items within each examination, it was necessary to assume that the ER items within each examination were somewhat similar to the ER items in the removed scholarship examinations. Admittedly, the scholarship examination items were considered more difficult than the provincial examination items. Although the difficulty may differ somewhat for these items, the format of the items was quite similar. Thus, for the purposes of the present study and given the ER format of the scholarship examinations, the ER items in the provincial examinations were used in the weighting procedures.

Fourth, due to the recent interest in the simultaneous estimation of examinations containing MC and ER items, the choice of available commercial software to complete such analysis was limited to that of PARSCALE 3.1. While the program is able to simultaneously estimate dichotomously- and polytomously-scored items using the two-parameter framework, dichotomously-scored items estimated using the three-parameter model must be estimated separately from the polytomously-scored items. Thus, it was not possible to assess the effect of including the third parameter for the dichotomously-scored items with simultaneous estimation of both the MC and ER items.

Conclusions and Implications for Practice

The results of the study have implications from both a policy and psychometric perspective. From a policy perspective, to the extent that the original procedure can be considered representative of the correct scholarship decisions and given the limitations of the study, the 10% error rate in the awarding of scholarships associated with the current procedure must be addressed. The results indicate that for the two years and 11 examinations analysed, one in 10 students, or approximately 500 students, would have been treated unfairly by the adoption of the current scholarship procedure. Further, the types of classification errors in the current procedure were more likely to be false negative decisions in which students would be unfairly denied scholarships. If such results are consistent over time, the present use of the current procedure has implications for the British Columbia academic scholarship program. First, the higher false negative rate reduces costs since fewer scholarships would be awarded in comparison to the original procedure. Second, different students are now being given scholarships than those that would be given scholarships if the original procedure was still being used.

Overall, a larger number of students will no longer receive scholarships because of the removal of the scholarship examinations.

Further, the current examination procedure appears to differentially effect males and females with a greater number of the errors occurring for males, both negative and positive. Thus, males have been more affected by the change in procedure than females. Such a problem, if consistent, needs to be addressed in order to try to reduce the discrepancy in error rates among males and females.

In order to address the errors associated with the current procedure, 17 alternative procedures were attempted that would potentially have lower error rates and would not require additional costs to be implemented by the Ministry of Education. It had previously been theorized that the use of response patterns within the framework of item response theory would be superior to the total test score approach (Wainer & Thissen, 1993; Samejima, 1996). Admittedly, the correlations among total test score and ability estimates had previously been shown to be high in examinations having dichotomously-scored items (Fan, 1998, Anderson, 1999). However, the use of the GPCM procedure, which is an IRT approach based on response vectors, did allow for different scholarship scores to be calculated for students having the same raw score. It was further hypothesized that these differences in response vectors would produce examination scholarship scores, classifications, and decisions that were closer to those produced by the original procedure than the scholarship scores and decisions produced by the current procedure.

Nonetheless, while differences between the current and the GPCM procedures did occur, the GPCM approach did not provide superior results in comparison to the current procedure. While there may be other benefits to the use of the GPCM procedure, these benefits were not found to extend to providing superior estimates of student ability. Thus, the differences in response vectors between students having the same raw score could not be related to differences in ability as defined by the scholarship scores obtained in the original procedure. Rather, the differences in response vectors for students having the same raw score were more likely due to random examination performance differences between students of the same overall ability with respect to the construct being measured. Hence, within the context of the scholarship program in British Columbia, the use of a total score method of reporting scores is comparable to the ability estimates derived from the GPCM. Due to the ease of use associated with a total score method, it remains the preferred choice for the purposes of scholarship score reporting.

Similar conclusions were reached with respect to each of the alternative procedures. Differences between the alternative and the current procedures were concluded to be due to random variations rather than systematic effects associated with the procedure employed. It was concluded that incorporating the GPCM, subtest weighting, and/or the use of school-based mark, would not improve upon the errors associated with the current procedure. However, as detailed below, the results of the examination of these alternative procedures did provide additional evidence both in support and rejection of previous research with regards to gender-by-item format interactions and relevant psychometric issues.

Differential effects for females and males. The results of previous research had indirectly suggested that the change in the scholarship procedure could have differential effects on the decision consistency for the scholarship decisions for females and males since differential performance has been reported across format (e.g., Bolger & Kellaghan, 1990; Garner & Engelhard, 1998; Henderson, 1999). In particular, males have been shown to have superior performance on MC items while females have been shown to have superior performance on ER items, especially for higher ability examinees. With respect to the present study, the removal of the scholarship examinations, which were made up entirely of ER items, increased the importance of the MC items within the current procedure. This was expected to increase the false negative rates for females and the false positive rates for males. However, the results of the gender analysis for the current procedure indicated that the false negative and false positive error rates were similar for females than males. It was also speculated that the alternative procedures that increased the weighting of the ER section or included SBM as another item would be more beneficial to females than males. Such benefits were not found in the approaches using increased weighting of the ER section and were not consistently found in the approaches that used SBM. Hence, in contrast to previous findings, higher ability females and males were not differentially affected by approaches that altered the relative contributions of the MC and ER sections (Bridgeman, 1989; Schmitt et al., 1991; DeMars, 1998).

Simultaneous vs. separate estimation of multiple-choice and extended-response items. The simultaneous estimation of MC and ER examination items using IRT models has been criticized both theoretically and psychometrically (Luecht & Miller, 1992; Luecht, 1994). Of particular concern is the lack of unidimensionality, a necessary assumption for the use of IRT models that has been shown to effect the accuracy of the estimation process if not met (Ackerman, 1989; Way, Ansley, & Forsyth, 1988; De

Ayala, 1994, 1995). Nonetheless, other research has suggested that the traits of both the MC and ER items are similar enough to be included together (e.g., Thissen et al., 1994; Ercikan et al., 1998). The analysis of the examinations within this study supported these latter findings since no systematic differences were found in either the parameter or ability estimates as related to the simultaneous or separate estimation of the MC and ER sections. If the MC and ER items within the different examinations analysed were measuring different traits, the similarity in the traits were so high that the unidimensional IRT models used did not produce substantively different estimates with respect to the simultaneous or separate estimation of the different item types. Hence, concerns about the simultaneous estimation of both MC and ER items seem to be largely unfounded for many course specific achievement examinations as represented in the present study. Although some problems did occur with the use of PARSCALE 3.1 and the three-parameter model (see below), once these problems have been resolved, the simultaneous estimation of MC and ER items, as is used by many current examinations, appears to be a viable method to obtain ability estimates and, if necessary, item parameter estimates.

The use of weighting. Given that the scholarship examinations consisted entirely of ER items, it was expected that the ER items within the provincial examinations would be better predictors of scholarship scores. Research using polytomous IRT models has previously shown that polytomously-scored items provide more information and thus better measurement for higher ability students than dichotomously-scored items (Donoghue, 1994; Wilson & Wang, 1995; Carlson, 1996). While the ER items themselves did provide more information and peak at a higher ability then the MC items, overall, the MC items generally provided more information. It appears that more ER items would be required before differentially weighting these items would have any substantial or predictable impact on the examination and total scholarship scores and decisions. Hence, within the limitations of the study, these apparent benefits of the ER items were not realized. Further, as first discussed by Wainer and Thissen (1993), the natural weighting process that is implicit with the use of the two-parameter model in combination with the GPCM is equal to or superior to any additional weighting of the MC and ER sections.

The use of auxiliary information in the form of school-based mark (SBM). Given that the provincial examinations varied in length from 70 to 120 marks and had between 37 and 64 items, it was expected that SBM would have a small but positive effect on the estimation of scholarship scores and decisions. This, in turn would provide closer results to the original procedure than those produced by the current procedure. With respect to IRT models, the use of auxiliary information has been shown to increase the apparent length of an examination (Mislevy, 1987). As with the total test score model, the relative advantage of the inclusion of auxiliary information is greater for examinations with a small number of items or having few examinees on which to complete the estimates (Mislevy, 1987). The inclusion of SBM had a marginal effect on the error and decision consistency rates in the present study but was not consistently superior to the current procedure. Further, the sizes of the effects were such that they could also be attributed to random variations. Thus, the use of SBM was not found to improve upon the current procedure.

Of more interest was the consistent, albeit small, effect the inclusion of SBM had on the false negative and false positive error rates. The inclusion of SBM tended to reduce the false negative rate and increase the false positive rate. A likely explanation is that the use of school-based marks mediated the differences in performance that individual students had between the provincial and the scholarship examinations in the original procedure. For example, some students may have done relatively poorly on a provincial examination as compared to the corresponding scholarship examination. Other students would have done relatively poorly on the scholarship examination. Such differences in performance would not be detected in the current procedure and would thus affect the decision consistency when both the current and original procedures were compared. For students who did relatively poorly on the provincial examinations there would be an increased likelihood of a false negative decision. In contrast, for those students who did relatively poorly on the scholarship examinations there would be an increased likelihood of a false positive decision. The use of SBM led to changes in the error rates consistent with what would be expected if performance between the two examinations was more consistent.

Lastly, the item parameter estimates for the SBM were found to vary depending on the approach used to obtain the estimates. For example *a*- and *b*-parameter estimates were very different when the SBM was estimated alone or simultaneously estimated with the MC and the ER items. Both increases and decreases in the *a*-parameter for the SBM were observed when the SBM was estimated with the examination items. It is possible that these differences are symptomatic of multidimensionality between the SBM and the corresponding examination scores.

The stability of the three-parameter model within PARSCALE 3.1. Unexpected problems using the three-parameter model within PARSCALE 3.1 prompted an exploratory analysis of the estimation procedure used in PARSCALE 3.1 as compared to BILOG 3.11. In the June 1995 Chemistry examination, in which convergence was achieved using the three-parameter model within the PARSCALE 3.1 program, the ability (θ) and *a*-parameter estimates were comparable to those obtained using BILOG 3.11 while the *b*- and *c*-parameter estimates were slightly less comparable having correlations of 0.91 and 0.88, respectively. Of greater concern was the number of examinations in which the estimation procedure within PARSCALE 3.1 did not achieve convergence and the extreme item parameter estimates produced for these examinations. A comparison of the June 1995 Geography dichotomously-scored items as estimated in PARSCALE 3.1 and BILOG 3.11 illustrated the extent of the problem. Unlike PARSCALE 3.1, the estimation process in BILOG 3.11 did achieve convergence and produced reasonable item parameter estimates. Further, the estimates produced by the two programs were less comparable than those obtained with the Chemistry examination. The correlations between the θ, *a*-, *b*-, and *c*-parameter estimates were 0.91, 0.75, 0.56, and 0.30, respectively. Based on these results, caution must be expressed about the use of the three-parameter model in PARSCALE 3.1.

Future Research

The results of the current study provide directions to pursue both from a policy perspective to address the problems associated with the current procedure as well as in terms of future research for the analysis of examinations having dichotomously- and polytomously-scored items. From the policy perspective, the current procedure has high false negative classification rates at the examination score level. This suggests that many of the students who do not obtain the 475 minimum examination scholarship score would obtain this minimum score if the original procedure was still being used. Since only those scores that are at least 475 count towards the total scholarship score, the current procedure eliminates a larger proportion of the students than the original procedure.

It is necessary for the British Columbia Ministry of Education to review the current procedure in order to reduce the scholarship error rate. One option would be to reintroduce the scholarship examinations in their original form. Although this would once again increase the costs of the examination program, it would eliminate the error rates found in the present study. A second option to be explored is to adjust the 70% minimum score downward. This would increase the number of students obtaining scholarship scores and also increase the individual student scholarship scores for those students having provincial examination scores of 70% or more. For example, a score of 75% could conceivably be the mean of those scores considered for scholarships and would thus be translated into a scholarship score of 500 based on the current procedure. If provincial examination scores of 69% were also included in the examination scholarship score calculation process, the mean score would be lowered and a student with 75% would receive a scholarship score higher than 500. An added benefit to this proposal is that it would also reduce the usage of the Kozlow correction formula, which was often required to adjust the scores within the current procedure. A related solution is to change the distributional characteristics of the scholarship scores, for example to a mean of 525 and a standard deviation of 100. A third option is to remove or reduce the 475 minimum while keeping the 1700 minimum total score. Students who do extremely well on two scholarship examinations but failed to meet the 475 criteria on a third examination would then have three examination scholarship scores in order to calculate a total scholarship score. It is likely that each of the procedures above would also increase the false positive rate. Preliminary analysis of the procedure eliminating the examination scholarship score minimum has indicated the decrease in the false negative rate would be larger than the corresponding increase in the false positive rate. Finally, scholarships could be given on a sliding scale. For example, students having a total scholarship score of 1600 to 1699 could be given a scholarship of $500.00 while those having a score of 1700 or more would continue to be given the $1000.00 scholarship currently offered. While this procedure would not eliminate the overall error rate, it would lessen the impact of such decision errors. Further, it would provide more money and opportunity to students rather than examination developers and markers.

The delimitations of the current study also provide avenues for further research. For example, the errors associated with those examinations having fewer than 1000 students were not included in the current study. The negative correlation between sample size and RMSE suggests that greater discrepancies exist between the original and the current procedures in those examinations that have fewer students. If it can be shown that students who enroll in the less "popular" provincially examinable courses have an unfair advantage or disadvantage in obtaining scholarships, procedures must be examined that will negate this difference.

Second, with the exception of Geography, the humanities examinations were not included in the current study due to the method in which the holistic scores were reported for the ER items. The examinations analysed in the current study were generally from the natural sciences and Mathematics and all used analytical scoring for the ER items. It is possible that differential effects would be observed in the humanities examinations not only because of differences in the scoring procedures used but also because of differences in the students that typically enroll in a predominantly humanities stream of education as opposed to a more science dominated stream. Further, the Geography examinations were the examinations that were least able to meet the assumptions of unidimensionality. Further research needs to be completed that explores the estimation procedures using actual data from examinations using holistic scales and those having varying degrees of dimensionality across the items. As echoed by Luecht (1994) it is reasonable to assume multidimensionality in examinations that measure writing skills and knowledge of grammar.

A second direction for research extending from this issue is the continued search to find methods to identify multidimensionality within achievement examinations. It appears that the present procedures are often insensitive to the detection of other dimensions. One possible avenue of research as suggested by the findings of this study is to actually complete the estimation process within a unidimensional IRT framework. The differences that occur in the item parameter estimates when estimation is done simultaneously or separately for those items predicted to be measuring a related but distinct trait would provide a measure of the presence of multidimensionality and its effect. Researchers have noted that differences occur in the parameter estimates when multidimensional data is used to obtain estimates in a unidimensional framework, but it has yet to be suggested or tested as a method to explore the assumption of unidimensionality (Yen, 1986; Luecht and Miller, 1992; De Ayala, 1994, 1995).

Finally, a direction for research as suggested by the results of the current study is an analysis of the relationship between the number of score categories, discrimination, and the amount of information an item provides. In the current study, the *a*-parameters for the polytomously-scored items were not only lower than the dichotomously-scored items but also lower than the *a*-parameters for polytomously-scored items as reported in other research (Donoghue, 1994; Fitzpatrick et al, 1996). Due to the inclusion of half-point marks, several of the polytomously-scored items analysed in the provincial examinations had 10 or more score categories. Although the GPCM as operationalized in PARSCALE 3.1 is able to estimate an item having 15 score categories, most previous research has used items with far fewer score categories, generally in the range of 4 to 6. Muraki (1993) has previously shown that the recoding (through rounding) of low response categories could potentially increase information. It is possible that the large number of score categories reduced the information that could have been obtained for the ER items. Thus, given the increasing use of ER items and polytomous item response models, further research needs to be completed that provides some guidelines regarding the relationship between the number of score categories within these items and the information provided based on the estimation procedure.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and non compensatory multidimensional items. Applied Psychological Measurement, 13, 113-127.

Adams, R. A., & Wilson, M. (1992, April). A random coefficients multinominal logit: Generalizing Rasch models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Anderson, J. O. (1999). Does complex analysis (IRT) pay any dividends in achievement testing? The Alberta Journal of Educational Research, XLV (4), 344-352.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm Psychometrika, 35, 179-197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431-444.

Bock, R.D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. Journal of Educational Measurement, 34, 197-211.

Bolger, N. & Kellaghan, T. (1990) Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27, 165-174.

Bridgeman, B. (1989). Comparative validity of multiple choice and free-response items on the advanced placement examination in biology (College Board Report No.

89-2). New York: College Entrance Examination Board. (ERIC Document Reproduction Service No. ED 308 228)

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. Journal of Educational Measurement, 29, 253-271.

Burton, N. M. (1996). How have changes in the SAT affected women's math scores? Educational Measurement: Issues and Practice, 15 (4), 5-9.

Carlson, J. E. (1996, April). Information provided by polytomous and dichotomous items on certain NAEP instruments. Paper presented at the annual meeting of the American Educational Research association, New York, NY.

Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983). Graphical methods for data analysis. Belmont CA: Wadsworth International Group; Boston: Duxbury Press.

Cheliminsky, E., & York, R. L. (1994). Educational testing: The Canadian experience with standards, examinations, and assessments. General Accounting Office Report PEMD-93-11. Gaithersburg: MD. GAO.

Council of Chief State School Officers State Education Assessment Center. (Dec 1998). Key state education policies on K-12 education. Washington, DC: CCSSO.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. Applied Psychological Measurement, 18, 155-170.

De Ayala, R. J. (1995). The influence of multidimensionality on estimation in the partial credit model. Educational and Psychological Measurement, 55, 407-422.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. Applied Measurement in Education, 11, 279-299.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. Journal of Educational Measurement, 31, 295-311.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. Journal of Educational Measurement, 35, 137-154.

Fan, X. (1998). Item response theory and classical test score theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58, 357-381.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. Journal of Educational Measurement, 33, 291-314.

Folske, J. C., Gessaroli, M. E., & Swanson, D. B. (1999, April). Assessing the utility of an IRT-based method for using collateral information to estimate subscores. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC.

Garner, M., & Englehard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. Applied Measurement in Education, 12, 29-51.

Gorsuch, R. L. (1983). Factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Norwell, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Henderson, D. L. (1999). Investigation of differential item functioning in exit examinations across item format and subject area. Unpublished doctoral dissertation, University of Alberta, Edmonton, Canada.

Huynh, H., & Ferrara, S. (1994). A comparison of equal percentile and partial credit equatings for performance-based assessments composed of free-response items. Journal of Educational Measurement, 31, 125-141.

Klinger, D. A., & Boughton, K. A. (2000, May). The accuracy of the generalized partial credit model and the graded response model in performance based assessments. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Edmonton, Alberta, Canada.

Lafleur, C. & Ireland, D. (1999). Canadian and provincial approaches to learning assessments and educational performance indicators. Technical report submitted to Commonwealth Caribbean Program, Americas Branch: The Canadian International Development Agency.

Lane, S., Wang, N., & Magon, M. (1996). Gender related differential item functioning on a middle-school Mathematics performance assessment. Educational Researcher, 15, 21-27, 31.

Lien, A. J. (1980). Measurement and evaluation of learning. Dubuque, IA: Wm. C. Brown.

Lord, F. M. (1952). A theory of test scores. Psychometric Monographs, No. 7.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13, 517-548.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison Wesley.

Luecht, R. M. (1994). Marginal maximum likelihood estimation of the generalized partial credit model using collateral information. Greensboro, NC: ERM Technical Report.

Luecht, R. M., & Miller, T. R. (1992, April). Consideration of multidimensionality in polytomous items response models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Maydeau-Olivares, A., Drasgow, F., & Mead, A. D. (1994) Distinguishing among parametric item response models for polychotomous ordered data. Applied Psychological Measurement, 18, 245-256.

Masters, G. N. (1982) A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Mislevy, R. J. (1987). Exploiting Auxiliary information about examinees in the estimation of item parameters. Applied Psychological Measurement, 11, 81-91.

Mislevy, R. J., & Bock, R. D. (1997). BILOG 3.11: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software, Inc.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. Applied Psychological Measurement, 17, 351-363.

Muraki, E. (1997). A generalized partial credit model. in W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 153-168). New York, NY: Springer.

Muraki, E., & Bock, R. D. (1997). PARSCALE 3.0: IRT item analysis and test scoring for rating-scale data [Computer program]. Chicago, IL: Scientific Software International, Inc.

Ndalichako, J. L. (1997). Comparison of number right, item response, and finite state approaches to scoring multiple-choice items. Unpublished doctoral dissertation, University of Alberta, Edmonton, Canada.

O'Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. Applied Measurement in Education, 11, 331-351.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207- 230.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.

Samejima, F. (1969). Estimation of latent trait using a response pattern of graded scores. Psychometrika Monograph Supplement, 34 (Monograph Number 17).

Samejima, F. (1972). A general model for free-response data. Psychometrika Monograph Supplement, 37 (Monograph Number 18).

Samejima. F. (1977). The use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

Samejima, F. (1996, April). Polychotomous responses and the test score. Paper presented at the annual meeting of the National Council of Measurement in Education, New York, NY.

Schmitt, A. P., Mazzeo, J., & Bleistein, C. (1991, April). Are gender differences between Advanced Placement multiple choice and constructed response sections a function of multiple choice DIF? Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet based tests. Journal of Educational Measurement, 25, 15-29.

S.P.S.S. for Windows 9.0 , Statistical package for the social sciences [Computer Software]. (1999). Chicago, IL: SPSS Inc.

Stiggins, R. J. (1997). Student-centered classroom assessment. Upper Saddle River: Merrill.

Stout, W. F. (1987). A nonparametric approach to assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39-49.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thissen, D., Wainer, H. & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests: An analysis of two tests. *Journal of Educational Measurement, 31,* 113-123.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.) *Applications of item response theory.* Vancouver, British Columbia, Canada: Educational Research Institute of British Columbia.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12,* 239-252.

Wiggins, G. (1993). *Assessing student performance.* San Francisco, CA: Jossey Bass.

Wilson, M., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. Applied Psychological Measurement, 19, 51-71.

Yamamoto, K., & Kulick, E. (1992, April). An information-based approach to maintaining content validity and determining the relative value of polytomous and dichotomous items. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. Journal of Educational Measurement, 23, 299-325.

Yen, W. M. (1992, April). Scaling performance assessments: Strategies for managing local item dependence. Invited address presented at the annual meeting of National Council on Measurement in Education, San Francisco, CA.

Appendix A

Comparison of the Examination Scholarship Results Between the Original and the Alternative Procedures

Table 19

Comparison of Examination Scholarship Results Between the Original and the GPCM Procedures

Subject	1994/1995					1995/1996				
	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.91	46.25	155 (9.8)	6 (0.4)	10.1	0.91	40.98	119 (6.4)	27 (1.4)	7.8
Biology (June)	0.91	42.67	248 (7.7)	55 (1.7)	9.4	0.91	41.15	202 (5.6)	89 (2.5)	8.1
Chemistry (January)	0.91	59.30	255 (14.)1	1 (0.1)	14.2	0.90	48.36	130 (7.1)	38 (2.1)	9.2
Chemistry (June)	0.92	43.83	308 (7.9)	66 (1.7)	9.6	0.91	45.19	389 (9.3)	66 (1.6)	10.9
Geography (January)	0.73	74.24	166 (19.4)	22 (2.6)	22.0	0.67	77.59	216 (20.5)	38 (3.6)	24.1
Geography (June)	0.74	66.59	253 (14.7)	61 (3.6)	18.3	0.74	67.93	227 (11.9)	68 (3.6)	15.5
Geology (June)	0.80	57.89	35 (11.4)	12 (3.9)	15.4	0.85	64.23	59 (19.2)	7 (2.3)	21.4
Mathematics (January)	0.89	52.68	269 (11.9)	13 (0.6)	12.5	0.86	56.03	402 (14.5)	28 (1.0)	15.6
Mathematics (June)	0.90	51.40	501 (9.5)	44 (0.8)	10.3	0.88	52.50	560 (10.7)	87 (1.7)	12.3
Physics (January)	0.87	65.61	129 (16.7)	6 (0.8)	17.5	0.88	56.41	128 (13.5)	9 (1.0)	14.5
Physics (June)	0.88	49.51	193 (7.4)	102 (3.9)	11.3	0.88	50.51	317 (11.1)	73 (2.6)	13.6

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 20

Comparison of Examination Scholarship Results Between the Original and the Classical, (Scholarship) Procedures

Subject	1994/1995					1995/1996				
	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.90	48.65	168 (10.6)	11 (0.7)	11.3	0.91	42.52	127 (6.8)	34 (1.8)	8.6
Biology (June)	0.90	45.78	257 (8.0)	61 (1.9)	9.9	0.91	40.53	204 (5.7)	85 (2.4)	8.0
Chemistry (January)	0.91	58.13	258 (14.3)	2 (0.1)	14.4	0.90	47.23	136 (7.4)	42 (2.3)	9.7
Chemistry (June)	0.91	45.69	291 (7.4)	70 (1.8)	9.2	0.91	45.39	403 (9.7)	71 (1.7)	11.4
Geography (January)	0.75	72.18	172 (20.1)	19 (2.2)	22.4	0.71	76.74	210 (20.0)	28 (2.7)	22.6
Geography (June)	0.75	67.20	243 (14.2)	56 (3.3)	17.4	0.77	62.70	216 (11.3)	69 (3.6)	15.0
Geology (June)	0.82	56.02	31 (10.1)	12 (3.9)	14.1	0.85	64.51	59 (19.2)	5 (1.6)	20.8
Mathematics (January)	0.90	52.04	268 (11.9)	9 (0.4)	12.3	0.87	54.70	401 (14.5)	36 (1.3)	15.8
Mathematics (June)	0.90	50.62	498 (9.4)	41 (0.8)	10.2	0.88	52.25	549 (10.5)	78 (1.5)	11.9
Physics (January)	0.85	67.62	127 (16.5)	10 (1.3)	17.7	0.87	57.41	126 (13.3)	9 (1.0)	14.3
Physics (June)	0.87	50.25	198 (7.6)	97 (3.7)	11.3	0.87	51.91	355 (12.4)	76 (2.7)	15.1

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 21

Comparison of Examination Scholarship Results Between the Original and the Classical, (Optimal) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.91	47.46	159 (10.0)	8 (0.5)	10.5	0.92	40.57	122 (6.5)	25 (1.3)	7.9
Biology (June)	0.92	42.39	243 (7.5)	51 (1.6)	9.1	0.92	39.19	200 (5.5)	80 (2.2)	7.8
Chemistry (January)	0.91	58.74	251 (13.9)	3 (0.2)	14.1	0.90	47.48	132 (7.2)	41 (2.2)	9.4
Chemistry (June)	0.92	44.16	293 (7.5)	69 (1.8)	9.3	0.92	45.01	394 (9.4)	68 (1.6)	11.1
Geography (January)	0.76	69.82	165 (19.3)	21 (2.5)	21.8	0.72	73.78	208 (19.8)	31 (2.9)	22.7
Geography (June)	0.76	65.16	245 (14.3)	56 (3.3)	17.5	0.78	61.98	212 (11.1)	62 (3.3)	14.4
Geology (June)	0.82	56.49	31 (10.1)	12 (3.9)	14.1	0.85	64.63	59 (19.2)	5 (1.6)	20.8
Mathematics (January)	0.90	52.12	265 (11.7)	14 (0.6)	12.4	0.87	55.43	416 (15.0)	34 (1.2)	16.3
Mathematics (June)	0.90	50.84	498 (9.4)	42 (0.8)	10.2	0.88	52.27	446 (8.5)	78 (1.5)	10.0
Physics (January)	0.88	64.18	132 (17.1)	9 (1.2)	18.3	0.89	55.02	121 (12.8)	7 (0.7)	13.5
Physics (June)	0.89	48.00	189 (7.3)	91 (3.5)	10.8	0.90	47.40	306 (10.7)	62 (2.2)	12.9

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 22

Comparison of Examination Scholarship Results Between the Original and the GPCM, (2-parm; Scholarship) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.90	46.78	163 (10.3)	10 (0.6)	10.9	0.91	42.30	124 (6.6)	32 (1.7)	8.4
Biology (June)	0.91	43.51	260 (8.1)	59 (1.8)	9.9	0.90	42.60	205 (5.7)	87 (2.4)	8.1
Chemistry (January)	0.88	61.41	256 (14.2)	3 (0.2)	14.4	0.90	48.07	141 (7.7)	46 (2.5)	10.2
Chemistry (June)	0.90	46.36	314 (8.0)	72 (1.8)	9.9	0.91	45.93	395 (9.5)	68 (1.6)	11.1
Geography (January)	0.75	71.23	163 (19.1)	20 (2.3)	21.4	0.69	76.52	210 (20.0)	34 (3.2)	23.2
Geography (June)	0.75	66.01	247 (14.4)	56 (3.3)	17.6	0.76	64.50	210 (11.0)	65 (3.4)	14.4
Geology (June)	0.80	57.28	35 (11.4)	12 (3.9)	15.4	0.85	65.00	59 (19.2)	7 (2.3)	21.4
Mathematics (January)	0.89	53.04	274 (12.1)	19 (0.8)	13.0	0.85	57.28	398 (14.4)	29 (1.0)	15.4
Mathematics (June)	0.89	51.86	525 (9.9)	56 (1.1)	11.0	0.87	54.13	561 (10.7)	81 (1.5)	12.2
Physics (January)	0.83	68.34	127 (16.5)	9 (1.2)	17.6	0.86	58.28	133 (14.1)	12 (1.3)	15.3
Physics (June)	0.87	51.21	192 (7.4)	96 (3.7)	11.1	0.86	53.12	327 (11.4)	78 2.7)	14.2

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 23

Comparison of Examination Scholarship Results Between the Original and the GPCM, (2-parm; Optimal) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.91	46.46	156 (9.8)	6 (0.4)	10.2	0.91	41.17	126 (6.7)	32 (1.7)	8.5
Biology (June)	0.92	42.06	249 (7.7)	56 (1.7)	9.5	0.90	42.23	204 (5.7)	88 (2.4)	8.1
Chemistry (January)	0.90	60.34	249 (13.8)	3 (0.2)	14.0	0.90	48.12	132 (7.2)	39 (2.1)	9.3
Chemistry (June)	0.92	44.21	308 (7.9)	60 (1.5)	9.4	0.91	45.34	391 (9.4)	68 (1.6)	11.0
Geography (January)	0.75	71.11	163 (19.1)	20 (2.3)	21.4	0.69	76.35	213 (20.2)	34 (3.2)	23.5
Geography (June)	0.76	64.95	246 (14.3)	54 (3.1)	17.5	0.76	64.70	212 (11.1)	65 (3.4)	14.5
Geology (June)	0.81	56.43	34 (11.1)	13 (4.2)	15.4	0.85	64.99	58 (18.8)	6 (1.9)	20.8
Mathematics (January)	0.89	53.03	274 (12.1)	19 (0.8)	13.0	0.86	56.40	399 (14.4)	29 (1.0)	15.5
Mathematics (June)	0.90	51.25	513 (9.7)	43 (0.8)	10.5	0.88	52.65	555 (10.6)	81 (1.5)	12.1
Physics (January)	0.87	64.71	128 (16.6)	6 (0.8)	17.4	0.88	56.40	129 (13.7)	8 (0.8)	14.5
Physics (June)	0.88	49.86	190 (7.3)	95 (3.7)	11.0	0.88	49.62	308 (10.8)	64 (2.2)	13.0

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 24

Comparison of Examination Scholarship Results Between the Original and the GPCM, (3-Parm; Scholarship) Procedures

Subject	1994/1995					1995/1996				
	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.90	47.26	160 (10.1)	8 (0.5)	10.6					
Biology (June)	0.90	43.76	257 (8.0)	54 (1.7)	9.6					
Chemistry (January)	0.89	60.56	257 (14.2)	3 (0.2)	14.4					
Chemistry (June)	0.90	47.03	323 (8.3)	79 (2.0)	10.3					
Geography (January)	0.75	70.65	164 (19.2)	21 (2.5)	21.7					
Geography (June)	0.74	68.57	252 (14.7)	61 (3.6)	18.2					
Geology (June)	0.80	57.82	35 (11.4)	12 (3.9)	15.4					
Mathematics (January)	0.90	52.56	272 (12.0)	22 (1.0)	13.0					
Mathematics (June)	0.89	52.07	534 (10.1)	57 (1.1)	11.2					
Physics (January)	0.83	68.36	127 (16.5)	9 (1.2)	17.6					
Physics (June)	0.87	50.54	194 (7.5)	97 (3.7)	11.2					

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 25

Comparison of Examination Scholarship Results Between the Original and the GPCM, (3-Parm; Optimal) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.91	47.21	157 (9.9)	8 (0.5)	10.4					
Biology (June)	0.91	42.78	250 (7.7)	50 (1.5)	9.3					
Chemistry (January)	0.91	59.65	255 (14.1)	1 (0.1)	14.2					
Chemistry (June)	0.92	44.41	309 (7.9)	63 (1.6)	9.5					
Geography (January)	0.75	70.60	166 (19.4)	22 (2.6)	22.0					
Geography (June)	0.74	68.59	251 (14.6)	61 (3.6)	18.2					
Geology (June)	0.81	56.44	34 (11.1)	13 (4.2)	15.4					
Mathematics (January)	0.90	52.29	271 (12.0)	17 (0.8)	12.7					
Mathematics (June)	0.90	50.93	508 (9.6)	46 (0.9)	10.5					
Physics (January)	0.87	65.25	130 (16.8)	7 (0.9)	17.7					
Physics (June)	0.88	49.55	194 (7.5)	97 (3.7)	11.2					

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 26

Comparison of Examination Scholarship Results Between the Original and the Classical, MC-ER-SBM Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.91	44.74	161 (10.1)	8 (0.5)	10.6	0.92	39.28	120 (6.4)	28 (1.5)	7.9
Biology (June)	0.91	41.32	238 (7.4)	58 (1.8)	9.2	0.92	37.66	173 (4.8)	103 (2.9)	7.7
Chemistry (January)	0.91	57.86	246 (13.6)	2 (0.1)	13.7	0.91	45.63	123 (6.7)	42 (2.3)	9.0
Chemistry (June)	0.92	43.50	259 (6.6)	71 (1.8)	8.4	0.92	44.68	400 (9.6)	72 (1.7)	11.3
Geography (January)	0.75	65.94	161 (18.9)	19 (2.2)	21.1	0.72	68.05	196 (18.6)	32 (3.0)	21.7
Geography (June)	0.76	61.38	221 (12.9)	68 (4.0)	16.8	0.78	59.55	201 (10.6)	76 (4.0)	14.5
Geology (June)	0.83	51.28	32 (10.5)	13 (4.2)	14.7	0.86	61.02	57 (18.5)	4 (1.3)	19.8
Mathematics (January)	0.90	50.02	257 (11.4)	12 (0.5)	11.9	0.88	53.24	412 (14.9)	27 (1.0)	15.9
Mathematics (June)	0.90	49.21	485 (9.2)	49 (0.9)	10.1	0.88	50.66	563 (10.7)	76 (1.4)	12.2
Physics (January)	0.88	61.51	127 (16.5)	7 (0.9)	17.4	0.88	55.56	123 (13.0)	6 (0.6)	13.7
Physics (June)	0.89	46.15	164 (6.3)	99 (3.8)	10.1	0.90	46.68	300 (10.5)	66 (2.3)	12.8

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 27

Comparison of Examination Scholarship Results Between the Original and the Classical, and SBM (Optimal) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.91	44.74	161 (10.1)	8 (0.5)	10.6	0.92	39.43	120 (6.4)	27 (1.4)	7.9
Biology (June)	0.91	41.39	234 (7.3)	58 (1.8)	9.0	0.92	37.74	171 (4.7)	103 (2.9)	7.6
Chemistry (January)	0.91	57.94	251 (13.9)	2 (0.1)	14.0	0.91	45.69	119 (6.5)	43 (2.3)	8.8
Chemistry (June)	0.92	43.50	257 (6.6)	72 (1.8)	8.4	0.92	44.68	375 (9.0)	74 (1.8)	10.8
Geography (January)	0.75	68.55	160 (18.7)	19 (2.2)	21.0	0.72	67.37	189 (18.0)	32 (3.0)	21.0
Geography (June)	0.76	62.39	257 (15.0)	58 (3.4)	18.3	0.78	60.90	220 (11.6)	65 (3.4)	15.0
Geology (June)	0.83	51.25	32 (10.5)	13 (4.2)	14.7	0.85	64.81	58 (18.8)	5 (1.6)	20.5
Mathematics (January)	0.90	50.33	264 (11.7)	10 (0.4)	12.1	0.88	53.25	408 (14.8)	29 (1.0)	15.8
Mathematics (June)	0.90	49.34	485 (9.2)	49 (0.9)	10.1	0.88	50.71	560 (10.7)	77 (1.5)	12.1
Physics (January)	0.88	63.13	131 (17.0)	7 (0.9)	17.9	0.88	55.57	122 (12.9)	7 (0.7)	13.7
Physics (June)	0.89	46.20	179 (6.9)	87 (3.3)	10.2	0.90	46.60	298 (10.4)	66 (2.3)	12.7

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 28

Comparison of Examination Scholarship Results Between the Original and the Classical, MC/ER/SBM (Optimal) Procedures

Subject	1994/1995					1995/1996				
	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.92	44.63	157 (9.9)	8 (0.5)	10.4	0.92	38.79	115 (6.2)	25 (1.3)	7.5
Biology (June)	0.92	40.60	237 (7.3)	57 (1.8)	9.1	0.92	37.80	184 (5.1)	87 (2.4)	7.5
Chemistry (January)	0.92	57.58	254 (14.1)	3 (0.2)	14.2	0.91	45.64	128 (7.0)	39 (2.1)	9.1
Chemistry (June)	0.92	43.01	285 (7.3)	68 (1.7)	9.0	0.92	43.83	386 (9.3)	68 (1.6)	10.9
Geography (January)	0.76	67.32	165 (19.3)	21 (2.5)	21.8	0.73	68.51	195 (18.5)	28 (2.7)	21.2
Geography (June)	0.77	62.39	231 (13.5)	57 (3.3)	16.8	0.79	60.10	205 (10.8)	69 (3.6)	14.4
Geology (June)	0.83	52.63	32 (10.5)	12 (3.9)	14.4	0.85	64.63	59 (19.2)	5 (1.6)	20.8
Mathematics (January)	0.90	50.63	263 (11.6)	10 (0.4)	12.1	0.88	53.03	400 (14.5)	31 (1.1)	15.6
Mathematics (June)	0.91	49.02	495 (9.4)	42 (0.8)	10.2	0.88	50.22	535 (10.2)	75 (1.4)	11.6
Physics (January)	0.88	61.99	132 (17.1)	8 (1.0)	18.1	0.89	54.49	119 (12.6)	8 (0.8)	13.4
Physics (June)	0.89	46.14	177 (6.8)	94 (3.6)	10.4	0.90	46.34	299 (10.5)	56 (2.0)	12.4

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 29

Comparison of Examination Scholarship Results Between the Original and the GPCM, MC-ER-SBM Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.92	42.85	152 (9.6)	8 (0.5)	10.1	0.92	39.26	120 (6.4)	31 (1.7)	8.1
Biology (June)	0.92	40.69	244 (7.6)	69 (2.1)	9.7	0.91	39.03	188 (5.2)	91 (2.5)	7.7
Chemistry (January)	0.90	57.70	241 (13.4)	1 (0.1)	13.4	0.90	46.12	126 (6.9)	40 (2.2)	9.1
Chemistry (June)	0.92	42.97	291 (7.4)	62 (1.6)	9.0	0.91	44.06	384 (9.2)	77 (1.8)	11.1
Geography (January)	0.65	79.62	174 (20.4)	25 (2.9)	23.3	0.69	72.38	201 (19.1)	36 (3.4)	22.5
Geography (June)	0.75	62.51	223 (13.0)	63 (3.7)	16.7	0.75	63.71	208 (10.9)	73 (3.8)	14.8
Geology (June)	0.81	54.77	33 (10.8)	15 (4.9)	15.7	0.85	61.77	56 (18.2)	5 (1.6)	19.8
Mathematics (January)	0.90	50.42	266 (11.8)	17 (0.8)	12.5	0.87	53.79	389 (14.1)	24 (0.9)	14.9
Mathematics (June)	0.90	49.58	502 (9.5)	56 (1.1)	10.6	0.88	50.34	532 (10.1)	85 (1.6)	11.7
Physics (January)	0.87	61.22	126 (16.3)	7 (0.9)	17.2	0.87	56.60	121 (12.8)	7 (0.7)	13.5
Physics (June)	0.88	47.57	174 (6.7)	99 (3.8)	10.5	0.88	48.82	304 (10.6)	72 (2.5)	13.1

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 30

Comparison of Examination Scholarship Results Between the Original and the GPCM, MC-ER/SBM (0.90, 0.1) Procedures

Subject	1994/1995					1995/1996				
	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.92	43.71	153 (9.6)	8 (0.5)	10.1	0.92	39.37	120 (6.4)	26 (1.4)	7.8
Biology (June)	0.92	41.06	250 (7.7)	63 (2.0)	9.7	0.91	39.28	192 (5.3)	90 (2.5)	7.8
Chemistry (January)	0.91	58.02	250 (13.9)	1 (0.1)	13.9	0.90	46.49	129 (7.0)	36 (2.0)	9.0
Chemistry (June)	0.92	42.87	299 (7.6)	62 1.6	9.2	0.92	44.16	386 (9.3)	67 (1.6)	10.9
Geography (January)	0.73	70.29	161 (18.9)	25 (2.9)	21.8	0.68	73.48	205 (19.5)	34 (3.2)	22.7
Geography (June)	0.75	63.80	238 (13.9)	63 (3.7)	17.5	0.75	64.54	211 (11.1)	69 (3.6)	14.7
Geology (June)	0.81	55.61	35 (11.4)	15 (4.9)	16.3	0.85	61.51	57 (18.5)	6 (1.9)	20.5
Mathematics (January)	0.90	51.02	263 (11.6)	16 (0.7)	12.4	0.87	54.53	392 (14.2)	22 (0.8	15.0
Mathematics (June)	0.90	49.89	501 (9.5)	47 (0.9)	10.4	0.88	50.85	549 (10.5)	85 (1.6)	12.1
Physics (January)	0.87	63.32	122 (15.8)	6 (0.8)	16.6	0.88	55.83	126 (13.3)	8 (0.8)	14.2
Physics (June)	0.88	48.01	182 (7.0)	96 (3.7)	10.7	0.88	49.52	312 (10.9)	67 (2.3)	13.3

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 31

Comparison of Examination Scholarship Results Between the Original and the GPCM, MC-ER/SBM (Optimal) Procedures

Subject	1994/1995					1995/1996				
	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.92	43.86	153 (9.6)	8 (0.5)	10.1	0.92	38.92	122 (6.5)	30 (1.6)	8.1
Biology (June)	0.92	40.73	248 (7.7)	66 (2.0)	9.7	0.91	38.72	187 (5.2)	93 (2.6)	7.8
Chemistry (January)	0.91	58.75	249 (13.8)	1 (0.1)	13.9	0.90	45.64	125 (6.8)	37 (2.0)	8.8
Chemistry (June)	0.92	42.66	294 (7.5)	61 (1.6)	9.1	0.91	43.87	383 (9.2)	78 (1.9)	11.1
Geography (January)	0.73	71.51	158 (18.5)	25 (2.9)	21.4	0.69	71.46	199 (18.9)	36 (3.4)	22.3
Geography (June)	0.75	63.86	237 (13.8)	63 (3.7)	17.5	0.75	64.87	214 (11.2)	68 (3.6)	14.8
Geology (June)	0.81	54.70	33 (10.8)	15 (4.9)	15.7	0.85	64.23	59 (19.2)	7 (2.3)	21.4
Mathematics (January)	0.90	50.57	263 (11.6)	15 (0.7)	12.3	0.87	53.86	390 (14.1)	23 (0.8)	14.9
Mathematics (June)	0.90	49.63	500 (9.5)	50 (0.9)	10.4	0.88	50.33	539 (10.3)	84 (1.6)	11.9
Physics (January)	0.87	62.67	124 (16.1)	6 (0.8)	16.8	0.88	55.53	124 (13.1)	8 (0.8)	14.0
Physics (June)	0.88	47.14	173 (6.7)	93 (3.6)	10.2	0.89	48.68	312 (10.9)	72 (2.5)	13.4

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 32

Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER/SBM (0.45, 0.45, 0.1) Procedures

Subject	1994/1995					1995/1996				
	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.92	43.86	153 (9.6)	7 (0.4)	10.1	0.92	39.57	123 (6.6)	33 (1.8)	8.4
Biology (June)	0.92	40.52	248 (7.7)	62 (1.9)	9.6	0.91	40.51	197 (5.5)	90 (2.5)	8.0
Chemistry (January)	0.90	58.57	248 (13.7)	1 (0.1)	13.8	0.90	46.09	127 (6.9)	39 (2.1)	9.1
Chemistry (June)	0.92	43.49	294 (7.5)	58 (1.5)	9.0	0.92	44.23	386 (9.3)	69 (1.7)	10.9
Geography (January)	0.72	70.80	160 (18.7)	27 (3.2)	21.9	0.67	74.91	208 (19.8)	37 (3.5)	23.3
Geography (June)	0.73	65.23	241 (14.0)	65 (3.8)	17.8	0.74	66.17	227 (11.9)	75 (3.9)	15.9
Geology (June)	0.81	54.25	34 (11.1)	14 (4.6)	15.7	0.85	61.54	56 (18.2)	5 (1.6)	19.8
Mathematics (January)	0.90	51.34	266 (11.8)	19 (0.8)	12.6	0.86	55.28	392 (14.2)	25 (0.9)	15.1
Mathematics (June)	0.90	49.98	514 (9.7)	52 (1.0)	10.7	0.88	51.78	547 (10.4)	82 (1.6)	12.0
Physics (January)	0.87	62.82	123 (15.9)	6 (0.8)	16.7	0.88	55.68	126 (13.3)	8 (0.8)	14.2
Physics (June)	0.88	48.44	183 (7.0)	96 (3.7)	10.7	0.89	48.78	305 (10.7)	61 (2.1)	12.8

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 33

Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER/SBM (Optimal) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.92	43.54	153 (9.6)	7 (0.4)	10.1	0.92	39.31	121 (6.5)	32 (1.7)	8.2
Biology (June)	0.92	40.30	249 (7.7)	63 (2.0)	9.7	0.91	39.90	192 (5.3)	89 (2.5)	7.8
Chemistry (January)	0.90	58.75	246 (13.6)	1 (0.1)	13.7	0.90	45.77	128 (7.0)	38 (2.1)	9.1
Chemistry (June)	0.92	43.05	303 (7.7)	64 (1.6)	9.4	0.91	43.99	381 (9.1)	74 (1.8)	10.9
Geography (January)	0.75	68.60	162 (19.0)	20 (2.3)	21.3	0.71	70.09	195 (18.5)	32 (3.0)	21.6
Geography (June)	0.76	62.20	236 (13.7)	57 (3.3)	17.1	0.77	61.98	206 (10.8)	69 (3.6)	14.4
Geology (June)	0.81	53.80	33 (10.8)	14 (4.6)	15.4	0.85	64.99	58 (18.8)	6 (1.9)	20.8
Mathematics (January)	0.90	51.24	265 (11.7)	18 (0.8)	12.5	0.87	54.16	393 (14.2)	27 (1.0)	15.2
Mathematics (June)	0.90	49.66	505 (9.6)	44 (0.8)	10.4	0.88	50.71	547 (10.4)	83 (1.6)	12.0
Physics (January)	0.88	62.45	126 (16.3)	6 (0.8)	17.1	0.88	55.47	124 (13.1)	8 (0.8)	14.0
Physics (June)	0.88	47.71	173 (6.7)	92 (3.5)	10.2	0.89	48.31	302 (10.6)	60 (2.1)	12.7

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 34

Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER-SBM (0.5, 0.5) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.92	43.85	155 (9.8)	8 (0.5)	10.3	0.91	39.80	123 (6.6)	35 (1.9)	8.5
Biology (June)	0.92	40.93	248 (7.7)	65 (2.0)	9.7	0.91	40.60	200 (5.5)	95 (2.6)	8.2
Chemistry (January)	0.90	58.06	243 (13.5)	2 (0.1)	13.6	0.90	46.03	128 (7.0)	38 (2.1)	9.1
Chemistry (June)	0.92	43.19	294 (7.5)	65 (1.7)	9.2	0.91	44.40	387 (9.3)	76 (1.8)	11.1
Geography (January)	0.71	71.50	161 (18.9)	29 (3.4)	22.2	0.66	76.01	207 (19.7)	38 (3.6)	23.3
Geography (June)	0.72	66.36	237 (13.8)	67 (3.9)	17.7	0.73	67.20	234 (12.3)	81 (4.3)	16.5
Geology (June)	0.81	54.06	34 (11.1)	15 (4.9)	16.0	0.85	62.32	57 (18.5)	6 (1.9)	20.5
Mathematics (January)	0.90	51.06	266 (11.8)	15 (0.7)	12.4	0.87	53.89	395 (14.3)	25 (0.9)	15.2
Mathematics (June)	0.90	49.52	511 (9.7)	58 (1.1)	10.8	0.88	50.55	544 (10.4)	85 (1.6)	12.0
Physics (January)	0.88	61.76	128 (16.6)	8 (1.0)	17.6	0.87	56.37	125 (13.2)	9 (1.0)	14.2
Physics (June)	0.88	48.37	177 (6.8)	98 (3.8)	10.6	0.89	48.30	299 (10.5)	59 (2.1)	12.5

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.

Table 35

Comparison of Examination Scholarship Results Between the Original and the GPCM; MC/ER-SBM (Optimal) Procedures

	1994/1995					1995/1996				
Subject	r	RMSE	False Negative	False Positive	Error Rate (%)	r	RMSE	False Negative	False Positive	Error Rate (%)
Biology (January)	0.91	43.54	154 (9.7)	8 (0.5)	10.2	0.91	40.56	122 (6.5)	33 (1.8)	8.3
Biology (June)	0.92	40.52	244 (7.6)	65 (2.0)	9.6	0.91	39.92	190 (5.3)	88 (2.4)	7.7
Chemistry (January)	0.90	58.80	243 (13.5)	1 (0.1)	13.5	0.90	46.17	128 (7.0)	38 (2.1)	9.1
Chemistry (June)	0.92	43.12	297 (7.6)	62 (1.6)	9.2	0.91	44.23	383 (9.2)	76 (1.8)	11.0
Geography (January)	0.75	65.64	149 (17.4)	21 (2.5)	19.9	0.71	70.23	196 (18.6)	33 (3.1)	21.8
Geography (June)	0.76	61.09	223 (13.0)	65 (3.8)	16.8	0.77	60.72	201 (10.6)	75 (3.9)	14.5
Geology (June)	0.81	54.18	34 (11.1)	15 (4.9)	16.0	0.85	61.83	56 (18.2)	5 (1.6)	19.8
Mathematics (January)	0.90	50.98	264 (11.7)	16 (0.7)	12.4	0.87	54.01	395 (14.3)	27 (1.0)	15.3
Mathematics (June)	0.90	49.60	505 (9.6)	57 (1.1)	10.6	0.88	50.43	541 (10.3)	87 (1.7)	12.0
Physics (January)	0.88	62.09	129 (16.7)	9 (1.2)	17.9	0.87	56.29	123 (13.0)	8 (0.8)	13.9
Physics (June)	0.88	48.68	176 (6.8)	97 (3.7)	10.5	0.89	48.28	302 (10.6)	61 (2.1)	12.7

Note: Differences in the overall error rate are due to rounding
Numbers in brackets represent the percentage of examinees in each category.